# The maximum entropy framework

## The maximum entropy principle — an example

Suppose we have a random variable $X$ with known states (values of the observations, $x_1, \ldots, x_n$) but unknown probabilities $p_1, \ldots, p_n$; plus some extra constrains, eg. $\langle X \rangle$ is known. We are given the task to attempt to have a good guess for the probabilities.

Example: $X$ can take 1, 2 or 3 with unknown probabilities, and $\langle X \rangle = \overline{x}$ is known. What is the "best guess" for the probabilities?

Need to find the maximum of $H(p_1, p_2, p_3)$ as a function of $p_1, p_2, p_3$, under two constraints: $\langle X \rangle = 1p_1 + 2p_2 + 3p_3 = \overline{x}$, and $p_1 + p_2 + p_3 = 1$.

$$\begin{aligned} 0 &= d\left[H(p_1,p_2,p_3) - \lambda\left(\sum_{i=1}^{3} ip_i - \overline{x}\right) - \mu\left(\sum_{i=1}^{3} p_i - 1\right)\right] \\ &= d\left[-\sum_{i=1}^{3} p_i \log p_i - \lambda \sum_{i=1}^{3} ip_i - \mu \sum_{i=1}^{3} p_i\right] \\ &= \sum_{i=1}^{3} \big\{ -\log p_i - 1 - \lambda i - \mu \big\} dp_i = 0 \end{aligned}$$

the curly brackets need to be zero:

$$-\log(p_i) - 1 - \lambda i - \mu = 0\,, \qquad\qquad i = 1,2,3$$

which with the notation $\lambda_0 = \mu + 1$ gives

$$p_i = e^{-\lambda_0 - \lambda i}\,.$$

The constraint on the sum of probabilities:

$$1 = \sum_{i=1}^{3} p_i = e^{-\lambda_0} \sum_{i=1}^{3} e^{-\lambda i} \qquad \Rightarrow \qquad e^{-\lambda_0} = \frac{1}{e^{-\lambda} + e^{-2\lambda} + e^{-3\lambda}}$$

so

$$p_i = \frac{e^{-\lambda i}}{e^{-\lambda} + e^{-2\lambda} + e^{-3\lambda}} = \frac{e^{\lambda(1-i)}}{1 + e^{-\lambda} + e^{-2\lambda}}$$

The other constraint, $\langle X \rangle = \overline{x}$:

$$\overline{x} = \sum_{i=1}^{3} ip_i = \frac{1 + 2e^{-\lambda} + 3e^{-2\lambda}}{1 + e^{-\lambda} + e^{-2\lambda}} \tag{1}$$

Multiplying the equation with the denominator gives a second degree equation for $e^{-\lambda}$:

$$(\overline{x} - 3)\left(e^{-\lambda}\right)^2 + (\overline{x} - 2)e^{-\lambda} + \overline{x} - 1 = 0$$

with this $p_2$ becomes

$$p_2 = \frac{-1 + \sqrt{4 - 3(\overline{x} - 2)^2}}{3}, \qquad p_1 = \frac{3 - \overline{x} - p_2}{2}\,, \qquad p_3 = \frac{\overline{x} - 1 - p_2}{2}$$

## Maximum entropy principle — general form

Suppose we have a random variable $X$ taking (known) values $x_1, \ldots, x_n$ with unknown probabilities $p_1, \ldots, p_n$. In addition, we have $m$ constraint functions $f_k(x)$ with $1 \le k \le m < n$, where

$$\langle f_k(X) \rangle = F_k\,,$$

the $F_k$s are fixed. Then the maximum entropy principle assigns probabilities in such a way that maximises the information entropy of X under the above constraints.

$$
\begin{aligned}
0 &= d\left[H(p_1,\ldots,p_n) - \sum_{k=1}^{m}\lambda_k\left(\sum_{i=1}^{n} f_k(x_i)p_i - F_k\right) - \underbrace{\mu}_{\lambda_0-1}\left(\sum_{i=1}^{n} p_i - 1\right)\right] \\
&= \sum_{i=1}^{n}\left\{-\log(p_i) - 1 - \sum_{k=1}^{m}\lambda_k f_k(x_i) - (\lambda_0 - 1)\right\} dp_i
\end{aligned}
$$

Since this is zero for any $dp_i$, all $n$ braces have to be zero:

$$
p_i = \exp\left(-\lambda_0 - \sum_{k=1}^{m}\lambda_k f_k(x_i)\right) \tag{2}
$$

The sum of probabilities give

$$
1 = \sum_{i=1}^{n} p_i = e^{-\lambda_0}\sum_{i=1}^{n}\exp\left(-\sum_{k=1}^{m}\lambda_k f_k(x_i)\right)
$$

Introducing *partition function*:

$$
Z(\lambda_1,\ldots,\lambda_m) := \sum_{i=1}^{n}\exp\left(-\sum_{k=1}^{m}\lambda_k f_k(x_i)\right) \tag{3}
$$

With this notation

$$
e^{-\lambda_0} = \frac{1}{Z(\lambda_1,\ldots,\lambda_m)}\,, \qquad \lambda_0 = \log Z(\lambda_1,\ldots,\lambda_m) \tag{4}
$$

The other constraints are

$$
\begin{aligned}
F_k &= \sum_{i=1}^{n} f_k(x_i)p_i = e^{-\lambda_0}\sum_{i=1}^{n} f_k(x_i)\exp\left(-\sum_{k=1}^{m}\lambda_k f_k(x_i)\right) = -\frac{1}{Z}\frac{\partial Z(\lambda_1,\ldots,\lambda_m)}{\partial\lambda_k} \\
&= -\frac{\partial\log Z(\lambda_1,\ldots,\lambda_m)}{\partial\lambda_k}\,,
\end{aligned} \tag{5}
$$

$$
p_i = \frac{1}{Z(\lambda_1,\ldots,\lambda_m)}\exp\left(-\sum_{k=1}^{m}\lambda_k f_k(x_i)\right) \tag{6}
$$

The value of the maximised information entropy:

$$
\begin{aligned}
S(F_1,\ldots,F_m) = H(\underbrace{p_1,\ldots,p_n}_{\text{from (6)}}) &= -\sum_{i=1}^{n} p_i\log(p_i) = -\sum_{i=1}^{n} p_i\left(-\lambda_0 - \sum_{k=1}^{m}\lambda_k f_k(x_i)\right) \\
&= \lambda_0 + \sum_{k=1}^{m}\lambda_k\sum_{i=1}^{n} f_k(x_i)p_i = \log Z(\lambda_1,\ldots,\lambda_m) + \sum_{k=1}^{m}\lambda_k F_k
\end{aligned} \tag{7}
$$

Now calculate the partial derivatives of $S$ w.r.t. the $F_k$s, being careful about what is kept constant in the partial derivatives

$$
\left.\frac{\partial S}{\partial F_k}\right|_{\{F\}} = \sum_{\ell=1}^{m}\underbrace{\left.\frac{\partial\log Z}{\partial\lambda_\ell}\right|_{\{\lambda\}}}_{F_\ell}\left.\frac{\partial\lambda_\ell}{\partial F_k}\right|_{\{F\}} + \sum_{\ell=1}^{m}\left.\frac{\partial\lambda_\ell}{\partial F_k}\right|_{\{F\}} F_\ell + \lambda_k = \lambda_k \tag{8}
$$