Negentropy. The negentropy of a random variable x is defined as :

$$J(x) = \hat{H}(x) - H(x)$$

where $H(x)$ is the differential entropy of x and $\hat{H}(x)$ the differential entropy considering a normal distribution with the same covariance matrix. Since the value of the differential entropy is maximal for a normal distribution, $J(x)$ is greater than 0. Furthermore, the more gaussian the distribution of x seems, the smaller the value of J.

Negentropy for clustering. Let $\omega = \{\omega_1, \cdots, \omega_k\}$ be a partition of a data space into k areas. The average negentropy of ω is calculated as follows :

$$\bar{J}(x) = \sum_{i=1}^{k} p_i J_i(x)$$

with $p_i = \frac{N_i}{N}$, N_i being the number of data belonging to cluster i and N the number of N.

Negentropy increment for clustering. To avoid complexity in calculus, rather than considering $\bar{J}$, we consider the negentropy increment which is the value of $\bar{J}$ less than the negentropy for the data considering that there is no partition J_0.

$$\Delta J = \bar{J}(x) - J_0(x) \quad (1)$$

$$= \sum_{i=1}^{k} p_i J_i(x) - (\hat{H}_0(x) - H_0(x)) \quad (2)$$

$$= \sum_{i=1}^{k} p_i J_i(x) - \hat{H}_0(x) + \sum_{x} p_0(x) ln(p_0(x)) \quad (3)$$

And because $p_0(x) = 1$ in the $H_0(x)$ calculation, each data belonging to the only one cluster existing (there is no partition considered), then we get :

$$\Delta J = \sum_{i=1}^{k} p_i J_i(x) - \hat{H}_0(x) \quad (4)$$

$$(5)$$

Then, we have $\hat{H}_0(x)$ the differential entropy of a normal distribution considering the covariance matrix of the oberved data which is given by Σ_0. This is by definition given by :

$$\frac{1}{2} ln((2\pi e)^N |\Sigma_0|) = \frac{N}{2} ln(2\pi e) + \frac{1}{2} ln(|\Sigma_0|)$$

We thus have :

$$\Delta J = \sum_{i=1}^{k} p_i J_i(x) - \frac{N}{2} ln(2\pi e) - \frac{1}{2} ln(|\Sigma_0|) \tag{6}$$

$$\Delta J = \sum_{i=1}^{k} p_i(\hat{H}_i(x) - H_i(x)) - \frac{N}{2} ln(2\pi e) - \frac{1}{2} ln(|\Sigma_0|) \tag{7}$$

We have $\hat{H}_i(x)$ the differential entropy of a normal distribution considering the covariance matrix of the oberved data which is given by Σ_i :

$$\hat{H}_i(x) = \frac{N_i}{2} ln(2\pi e) + \frac{1}{2} ln(|\Sigma_i|)$$

and

$$H_i(x) = -\sum_{x} p_i(x) ln(p_i(x))$$

We then get :

$$\Delta J = \sum_{i=1}^{k} \frac{N_i}{N}_i(x) - \frac{N}{2} ln(2\pi e) - \frac{1}{2} ln(|\Sigma_0|) \tag{8}$$

$$= \sum_{i=1}^{k} p_i(\frac{N_i}{2} ln(2\pi e) + \frac{1}{2} ln(|\Sigma_i|) + \sum_{x} p_i(x) ln(p_i(x))) - \frac{N}{2} ln(2\pi e) - \frac{1}{2} ln(|\Sigma_0|) \tag{9}$$

$$= \frac{1}{2} \sum_{i=1}^{k} (p_i N_i ln(2\pi e)) - \frac{N}{2} ln(2\pi e) + \sum_{i=1}^{k} p_i(\frac{1}{2} ln(|\Sigma_i|) + \sum_{x} p_i(x) ln(p_i(x))) - \frac{1}{2} ln(|\Sigma_0|) \tag{10}$$

Let

$$A = \frac{1}{2} \sum_{i=1}^{k} (p_i N_i ln(2\pi e)) - \frac{N}{2} ln(2\pi e)$$

, then :

$$\Delta J = A + \sum_{i=1}^{k} p_i(\frac{1}{2} ln(|\Sigma_i|) + \sum_{x} p_i(x) ln(p_i(x))) - \frac{1}{2} ln(|\Sigma_0|) \tag{11}$$

$$= A + \frac{1}{2} \sum_{i=1}^{k} p_i ln(|\Sigma_i|) + \sum_{i=1}^{k} p_i \sum_{x} p_i(x) ln(p_i(x))) - \frac{1}{2} ln(|\Sigma_0|) \tag{12}$$