Mathematical discussion of Equations (6) and (7)

Below, we explain in more detail how Equations (6) and (7) are constructed:

Distribution of r in the cluster p(r|C) (Equation (6))

Assume that all the gene expression vectors g_i are normalised (see normalisation section) and therefore are located in an E-dimensional space on the intersection of a hypersphere (with a radius equal to $\sqrt{(E-1)}$ (Equation (2))) and a hyperplane (Equation (1)) going through the center of the hypersphere. The intersection itself (we will further refer to it as H) can therefore be seen as a curved space with an intrinsic dimensionality of D=E-2 (H itself is a hypersphere with radius $\sqrt{(E-1)}$ located in the (E-1)-dimensional space defined by the hyperplane). We simplify the problem by neglecting the curved nature of H in the neighbourhood of the cluster (we assume the hypersphere to be locally flat - said otherwise, we linearise H in the neighbourhood of the cluster – we will refer to this linearised version of H as H_L). This approximation also implies that the cluster center C_K belongs to H_L and that the Euclidean distances to the cluster center measured in H_L are equal to the real Euclidean distances (=r) to the cluster center. The equations derived in this section are therefore an approximation and thus only reliable close to the current cluster center C_K ($r < \sqrt{(E-1)}$ = radius H) which is sufficient for our purpose, because we are only interested in modelling the area where the cluster is situated.

The cluster is assumed to be normally distributed around C_K within H_L (the variance is hypothesised to be equal in each direction (in H_L) and given by σ^2). This means that the probability of finding an expression vector g of the cluster in an elementary volume dV of H_L is given by (Bishop, 1995):

$$\frac{1}{\left(2\pi\sigma^2\right)^{D/2}}\exp\left(-\frac{\left\|g-C_K\right\|^2}{2\sigma^2}\right)dV = \frac{1}{\left(2\pi\sigma^2\right)^{D/2}}\exp\left(-\frac{r^2}{2\sigma^2}\right)dV, \quad (13)$$

where r is the Euclidean distance from the expression vector g to the cluster center C_K.

We know that the volume inside a shell with radius r around C_K in H_L (with elementary thickness dr) equals (Bishop, 1995)

$$\frac{2\pi^{D/2}}{\Gamma(D/2)}r^{D-1}dr = S_D r^{D-1}dr, \quad (14)$$

where S_D is the surface area of a unit sphere in D dimensions and Γ is the gamma function.

Replacing dV in Equation (13) by Equation (14) gives us the probability of finding an expression vector of the cluster inside the elementary shell:

$$\frac{S_D}{\left(2\pi\sigma^2\right)^{D/2}} r^{D-1} \exp\left(-\frac{r^2}{2\sigma^2}\right) dr = p(r \mid C)dr. \quad (15)$$

Said otherwise, Equation (15) results in the probability density estimation (p(r|C)) describing the distribution of r in the current cluster.

Distribution of r in the background p(r|B) (Equation (7))

As previously mentioned, H can be described as a D-dimensional curved space (hypersphere with radius $\sqrt{(E-1)}=\sqrt{(D+1)}$). It has a finite volume given by (Bishop, 1995):

$$S_{D+1}\left(D+1\right)^{D/2}, \quad (16)$$

where S_{D+1} is the surface area of a unit sphere in D+1 dimensions.

The background is assumed to be uniformly distributed in this finite volume. Dividing Equation (14) by Equation (16) gives us the probability of finding an expression vector of the background inside the elementary shell:

$$\frac{S_D}{S_{D+1}(D+1)^{D/2}} r^{D-1} dr = p(r \mid B)dr. \quad (17)$$

Said otherwise, Equation (17) results in the probability density estimation (p(r|B)) describing the distribution of r in the background.