Clustering

Andrew W. Rose

May 12, 2023

1 Introduction

The probability (density?) of a point with index, i, originating from a circular (spherical) Gaussian distribution with parameters (μ_d, σ) is

$$p(V_i \mid \mu, \sigma) = \prod_d \frac{1}{\sqrt{2\pi\overline{\sigma}_i^2}} \exp\left(\frac{-(x_{d,i} - \mu_d)^2}{2\overline{\sigma}_i^2}\right) \tag{1}$$

where d are the indices of the spacial dimensions, x_d are the spatial coordinates of the point, and $\overline{\sigma}_i^2 = \sigma^2 + s_i^2$.

If many points originate from the same given distribution, we may multiply the probabilities of the individual points. The probability that all the points belong to any spherical Gaussian distribution may be obtained by integrating over the parameters of the Gaussian

$$p(\nu) = \int_\sigma p(\sigma) \int_\mu p(\mu) \prod_{i=1}^{N} p(V_i \mid \mu, \sigma)\, d\mu\, d\sigma \tag{2}$$

Assuming a flat distribution of μ_d gives

$$p(\mu) = \prod_d (x_d^+ - x_d^-)^{-1} \tag{3}$$

$$= V^{-1} \tag{4}$$

Where V is the area (volume) over which we are integrating.

Since the circular (spherical) Gaussian is separable, we may consider the each integral over the spatial dimension of μ independently, and dropping the subscript 'd' for clarity:

$$\begin{aligned}
&\int_{x_d-}^{x_d+} \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi\overline{\sigma}_i^2}} \exp\left(\frac{-(x_i-\mu)^2}{2\overline{\sigma}_i^2} \right) \right) d\mu \\
&\qquad = \int_{x_d-}^{x_d+} \left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\overline{\sigma}_i^2}} \right) \left(\prod_{i=1}^{N} \exp\left(\frac{-(x_i-\mu)^2}{2\overline{\sigma}_i^2} \right) \right) d\mu && (5) \\
&\qquad = \left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\overline{\sigma}_i^2}} \right) \int_{x_d-}^{x_d+} \prod_{i=1}^{N} \exp\left(\frac{-(x_i-\mu)^2}{2\overline{\sigma}_i^2} \right) d\mu && (6) \\
&\qquad = \left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\overline{\sigma}_i^2}} \right) \int_{x_d-}^{x_d+} \exp\left(-\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\overline{\sigma}_i^2} \right) d\mu && (7)
\end{aligned}$$

Expanding the sum inside the integral and regrouping common terms in μ,

$$\begin{aligned}
\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\overline{\sigma}_i^2} &= \frac{1}{2} \left(\mu^2 \sum_{i=1}^{N} \frac{1}{\overline{\sigma}_i^2} - 2\mu \sum_{i=1}^{N} \frac{x_i}{\overline{\sigma}_i^2} + \sum_{i=1}^{N} \frac{x_i^2}{\overline{\sigma}_i^2} \right) && (8) \\
&= \frac{1}{2} \left(A\mu^2 - 2B\mu + C \right) && (9) \\
&= \frac{1}{2} \left(A(\mu - D)^2 + E \right) && (10)
\end{aligned}$$

Where (readding the subscript 'd'),

$$\begin{aligned}
A &= \sum_{i=1}^{N} \frac{1}{\overline{\sigma}_i^2} && (11) \\
B_d &= \sum_{i=1}^{N} \frac{x_{d,i}}{\overline{\sigma}_i^2} = AD_d && (12) \\
C_d &= \sum_{i=1}^{N} \frac{x_{d,i}^2}{\overline{\sigma}_i^2} = AD_d^2 + E_d = \frac{B_d^2}{A} + E_d && (13) \\
D_d &= \frac{B_d}{A} && (14) \\
E_d &= C_d - AD_d^2 = C_d - \frac{B_d^2}{A} && (15)
\end{aligned}$$

The form in equation 10 is convenient for taking the integral

$$\int_{x^-}^{x^+} \exp\left(-\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\overline{\sigma}_i^2}\right) d\mu$$

$$= \int_{x^-}^{x^+} \exp\left(-\frac{1}{2}\left(A(\mu-D)^2+E\right)\right) d\mu \tag{16}$$

$$= \int_{x^-}^{x^+} \exp\left(-\frac{E}{2}\right)\exp\left(-\frac{1}{2}\left(A(\mu-D)^2\right)\right) d\mu \tag{17}$$

$$= \exp\left(-\frac{E}{2}\right)\int_{x^-}^{x^+} \exp\left(-\frac{1}{2}\left(A(\mu-D)^2\right)\right) d\mu \tag{18}$$

$$= \exp\left(-\frac{E}{2}\right)\int_{x^-}^{x^+} \exp\left(\frac{-(\mu-D)^2}{2\frac{1}{A}}\right) d\mu \tag{19}$$

$$= \exp\left(-\frac{E}{2}\right)\int_{x^-}^{x^+} \frac{\sqrt{2\pi\frac{1}{A}}}{\sqrt{2\pi\frac{1}{A}}}\exp\left(\frac{-(\mu-D)^2}{2\frac{1}{A}}\right) d\mu \tag{20}$$

$$= \sqrt{\frac{2\pi}{A}}\cdot\exp\left(-\frac{E}{2}\right)\int_{x^-}^{x^+} \frac{1}{\sqrt{2\pi\frac{1}{A}}}\exp\left(\frac{-(\mu-D)^2}{2\frac{1}{A}}\right) d\mu \tag{21}$$

$$= \sqrt{\frac{2\pi}{A}}\cdot\exp\left(-\frac{E}{2}\right)\left[\phi\left(\frac{x^+-D}{\sqrt{\frac{1}{A}}}\right)-\phi\left(\frac{x^--D}{\sqrt{\frac{1}{A}}}\right)\right] \tag{22}$$

$$= \sqrt{\frac{2\pi}{A}}\cdot\exp\left(-\frac{E}{2}\right)\left[\phi\left(\sqrt{A}(x^+-D)\right)-\phi\left(\sqrt{A}(x^--D)\right)\right] \tag{23}$$

$$= \sqrt{\frac{2\pi}{A}}\cdot\exp\left(-\frac{E}{2}\right)\cdot G \tag{24}$$

Where we have defined

$$G_d = \phi\left(\sqrt{A}(x_d^+-D_d)\right)-\phi\left(\sqrt{A}(x_d^--D_d)\right) \tag{25}$$

We may, therefore, write

$$
\begin{aligned}
p(\nu) &= \int\limits_{\sigma} p(\sigma) \int\limits_{\mu} p(\mu) \prod_{i=1}^{N} p(V_i \mid \mu, \sigma)\, d\mu\, d\sigma \\
&= \int\limits_{\sigma} p(\sigma) \prod_{d} \left[\frac{1}{(x_d^+ - x_d^-)} \left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi \overline{\sigma}_i^2}} \right) \sqrt{\frac{2\pi}{A}} \exp\left(-\frac{E_d}{2}\right) G_d \right] d\sigma \qquad (26) \\
&= (2\pi)^{(1-N)\cdot\tau/2} \cdot V^{-1} \int\limits_{\sigma} p(\sigma) \cdot \left(\prod_{i=1}^{N} \frac{1}{\overline{\sigma}_i^2} \right)^{\tau/2} \cdot A^{-\tau/2} \cdot \prod_{d} \left[\exp\left(-\frac{E_d}{2}\right) \right] \cdot \prod_{d} G_d\, d\sigma \qquad (27) \\
&= (2\pi)^{(1-N)\cdot\tau/2} \cdot V^{-1} \int\limits_{\sigma} p(\sigma) \cdot F^{\tau/2} \cdot A^{-\tau/2} \cdot \exp\left(-\frac{1}{2} \sum_{d} E_d \right) \cdot \prod_{d} G_d\, d\sigma \qquad (28) \\
&= (2\pi)^{(1-N)\cdot\tau/2} \cdot V^{-1} \int\limits_{\sigma} p(\sigma) \cdot \left(\frac{F}{A} \right)^{\tau/2} \cdot \exp\left(-\frac{E}{2}\right) \cdot \prod_{d} G_d\, d\sigma \qquad (29)
\end{aligned}
$$

Where

$$
\begin{aligned}
E &= \sum_{d} E_d && (30) \\
&= \sum_{d} (C_d - B_d D_d) && (31) \\
&= \sum_{d} C_d - \sum_{d} B_d D_d && (32) \\
&= C - \sum_{d} B_d D_d && (33) \\
C &= \sum_{d} C_d && (34) \\
&= \sum_{d} \sum_{i=1}^{N} \frac{x_{d,i}^2}{\overline{\sigma}_i^2} && (35) \\
&= \sum_{i=1}^{N} \frac{\sum_d x_{d,i}^2}{\overline{\sigma}_i^2} && (36) \\
&= \sum_{i=1}^{N} \frac{r_i^2}{\overline{\sigma}_i^2} && (37) \\
F &= \prod_{i=1}^{N} \frac{1}{\overline{\sigma}_i^2} && (38)
\end{aligned}
$$

Noting that since the variable D_d can be expressed as a function of B_d, and that all the C_d are folded into C, equation 29 can be written in a form depending only on the "fundamental" variables:

$$A = \sum_{i=1}^{N} \frac{1}{\overline{\sigma}_i^2} \tag{39}$$

$$B_d = \sum_{i=1}^{N} \frac{x_{d,i}}{\overline{\sigma}_i^2} \tag{40}$$

$$C = \sum_{i=1}^{N} \frac{r_i^2}{\overline{\sigma}_i^2} \tag{41}$$

$$F = \prod_{i=1}^{N} \frac{1}{\overline{\sigma}_i^2} \tag{42}$$

If the product inside the integral is numerically unstable, so can instead perform the sum of logarithms and take the exponent only prior to performing the integral.

$$\begin{aligned} &ln\left(p(\sigma)\cdot\left(\frac{F}{A}\right)^{\tau/2}\cdot\exp\left(-\frac{E}{2}\right)\cdot\prod_d G_d\right) \\ &= ln\left(p(\sigma)\right)+\frac{\tau}{2}\left(ln\left(F\right)-ln\left(A\right)\right)-\frac{1}{2}E+\sum_d ln\left(G_d\right) \qquad (43) \\ & \qquad (44) \end{aligned}$$

And, we may simplify $ln\left(F\right)$ as

$$ln\left(F\right) = ln\left(\prod_{i=1}^{N} \frac{1}{\overline{\sigma}_i^2}\right) \tag{45}$$

$$= \sum_{i=1}^{N} ln\left(\frac{1}{\overline{\sigma}_i^2}\right) \tag{46}$$

$$p(l) = p_B^{n_B} \left(1 - p_B\right)^{N - n_B} \frac{\alpha^m \Gamma(\alpha) \prod_{k=1}^{m} \Gamma(n_k)}{\Gamma(\alpha + N - n_B)} \tag{47}$$

$$\begin{aligned}
ln\left(p\left(l\right)\right) &= n_B \cdot ln\left(p_B\right) + \left(N - n_B\right) \cdot ln\left(1 - p_B\right) + m \cdot ln\left(\alpha\right) \\
&\quad + ln\left(\Gamma\left(\alpha\right)\right) + ln\left(\prod_{k=1}^{m} \Gamma\left(n_k\right)\right) - ln\left(\Gamma\left(\alpha + N - n_B\right)\right) \qquad (48) \\
&= n_B \cdot ln\left(p_B\right) + \left(N - n_B\right) \cdot ln\left(1 - p_B\right) + m \cdot ln\left(\alpha\right) \\
&\quad + ln\left(\Gamma\left(\alpha\right)\right) + \sum_{k=1}^{m} ln\left(\Gamma\left(n_k\right)\right) - ln\left(\Gamma\left(\alpha + N - n_B\right)\right) \qquad (49) \\
&= n_B \cdot ln\left(p_B\right) + \left(N - n_B\right) \cdot ln\left(1 - p_B\right) + m \cdot ln\left(\alpha\right) \\
&\quad + \overline{\Gamma}\left(\alpha\right) + \sum_{k=1}^{m} \overline{\Gamma}\left(n_k\right) - \overline{\Gamma}\left(\alpha + N - n_B\right) \qquad (50)
\end{aligned}$$

where $\overline{\Gamma}$ is the ln of the Γ function.