The mathematics of Zalando's Next Top Analyst problem

Miłosz Jerkiewicz

This document contains some notes and mathematical work necessary to compute the probability distributions featuring in the problem.

1. Coordinate projection

The goal of the problem is to find the geographical coordinates (latitude, longitude) of the high-probability areas and present them in the form of an easy to read map. Because the probability distributions used to evaluate (latitude, longitude) points have their parameters specified in meters and they assume an orthogonal coordinate system, it is necessary to use a projection from one system to another. The following projection is provided in the problem statement:

$$x(\text{lat}, \text{lon}) = 111.323\ (\text{lon} - \text{sw\_lon})\ \cos\Big(\text{sw\_lat}\ \frac{\pi}{180}\Big)\ [km]$$
$$y(\text{lat}, \text{lon}) = 111.323\ (\text{lat} - \text{sw\_lat})\ [km]$$

Where

$$\text{sw\_lat} = 52.464011$$
$$\text{sw\_lon} = 13.274099$$

are the coordinates of the south-west point of the area of interest.

This projection returns (X, Y) coordinates specified in kilometers. This is a slight inconvenience, because the probability distributions have parameters specified im meters. By multiplying both equations by 1000 we get the final, ready-to-use equations:

$$x(\text{lat}, \text{lon}) = 111323\ (\text{lon} - \text{sw\_lon})\ \cos\Big(\text{sw\_lat}\ \frac{\pi}{180}\Big)\ [m]$$
$$y(\text{lat}, \text{lon}) = 111323\ (\text{lat} - \text{sw\_lat})\ [m]$$

2. Log-normal distribution

The probability density function of the log-normal distribution is given by the following formula:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$$

To compute $f(x)$ at a given point x, we need to know μ and σ. To compute them, we can use the following information:

> *The distribution's radial profile is log-normal with a mean of 4700m and a mode of 3877m in every direction.*

The starting point are the formulas for the mean and the mode of the log-normal distribution:

$$\text{mean} = e^{\,\mu+\frac{\sigma^2}{2}} \tag{1}$$

$$\text{mode} = e^{\,\mu-\sigma^2} \tag{2}$$

We need to solve this set of equations to get μ and σ. Logarithming both sides gives:

$$\mu + \frac{\sigma^2}{2} = \ln(\text{mean}) \tag{1'}$$

$$\mu - \sigma^2 = \ln(\text{mode}) \tag{2'}$$

Substracting (2') from (1')

$$\mu + \frac{\sigma^2}{2} - (\mu - \sigma^2) = \ln(\text{mean}) - \ln(\text{mode})$$

$$\frac{3}{2}\sigma^2 = \ln(\text{mean}) - \ln(\text{mode})$$

And finally:

$$\sigma = \sqrt{\frac{2}{3}\big(\ln(\text{mean}) - \ln(\text{mode})\big)} \tag{σ}$$

Getting back to equation (2'):

$$\mu - \sigma^2 = \ln(\text{mode}) \tag{2'}$$

$$\mu = \ln(\text{mode}) + \sigma^2$$

Plugging the σ computed above:

$$\mu = \ln(\text{mode}) + \frac{2}{3}\left(\ln(\text{mean}) - \ln(\text{mode})\right)$$
$$\mu = \ln(\text{mode}) + \frac{2}{3}\ln(\text{mean}) - \frac{2}{3}\ln(\text{mode})$$
$$\mu = \frac{2}{3}\ln(\text{mean}) + \frac{1}{3}\ln(\text{mode})$$

And finally:

$$\mu = \frac{2\ln(\text{mean}) + \ln(\text{mode})}{3} \qquad (\mu)$$

The end result is the following set of formulas for μ and σ:

$$\mu = \frac{2\ln(\text{mean}) + \ln(\text{mode})}{3} \qquad (\mu)$$
$$\sigma = \sqrt{\frac{2}{3}\left(\ln(\text{mean}) - \ln(\text{mode})\right)} \qquad (\sigma)$$

For the given data (*a mean of 4700m and a mode of 3877m*) the values are:

$$\mu = 8.39115083$$
$$\sigma = 0.358237212$$

3. Normal distribution

The probability density function of the normal distribution is given by the following formula:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

As it was the case with the log-normal distribution, we need to find μ and σ. We have the following quotes (two different distributions):

> *with 95% probability she is located within 2400m distance of the satellite's path*

> *The probability at any point is given by a Gaussian function of its shortest distance to the river. The function peaks at zero and has 95% of its total integral within ±2730m*

In each case, μ is 0 (because the probability density function peaks if the distance to the river or the satellite path is 0). What is left is to find the σ.

If 95% of the probability lies in range $< -x, x >$ (x is expressed in terms of σ, so 2 means 2σ etc.), then the remaining 5% is symetrically distributed in the ranges $(-\infty, x)$ and (x, ∞), 2.5% in each. This means that for x, the cumulative distribution function reaches 97.5%:

$$\text{cdf}(x) = 0.975$$

To find the value of x we take the inverse of cdf, called the *probit* function. The above equation can be rewritten as

$$x = \text{probit}(0.975)$$

which is approximately 1.95996398. The interpretation: the normal distribution places 95% of probability within the range $-1.95996398\,\sigma$ and $1.95996398\,\sigma$. The last step is to use the given 95% probability boundaries (*2400m* and *2730m*) to obtain σ:

$$\sigma = \frac{\text{95\% probability boundary}}{1.95996398}$$

4. Combining probability distributions

Consider the following problem: we are given three probability distributions that describe the random variable X. The distributions come from independent, reliable observers A, B and C. The question is, what is the true distribution of X? In other words, how can we combine the measured distributions to obtain the best approximation of the distribution of X?

What we are looking for is the *joint distribution* $P(x_A, x_B, x_C)$. It is the probability that A observes x_A while at the same time B observes x_B and C observes x_C. The observers make their observations independently, therefore $P(x_A, x_B, x_C) = P(x_A)P(x_B)P(x_C)$. This means that the observed probability distributions can be combined into one by multiplying them together.

5. Distance between two points (2 dimensions)

The distance between points P and Q is the length of the vector $Q - P$:

$$|Q - P| = \sqrt{(q_x - p_x)^2 + (q_y - p_y)^2}$$

It is used to compute the log-normal distribution centered in the Brandenburg Gate.

6. Nearest point on a line (in 2 dimensions)

We are given a line $\overleftrightarrow{KL}$ defined by points K and L. We are also given a point P. The objective is to find the point N on line $\overleftrightarrow{KL}$ that is nearest to the point P.

The point N lies on the intersection of the line $\overleftrightarrow{KL}$ and the line $\overleftrightarrow{PN}$ which has to be normal to $\overleftrightarrow{KL}$. The direction of the line $\overleftrightarrow{KL}$ is given by the vector

$$\mathbf{v} = L - K = (l_x - k_x, l_y - k_y)$$

The vector normal to $\mathbf{v}$ is

$$\mathbf{n} = (l_y - k_y, -(l_x - k_x)) = (l_y - k_y, k_x - l_x)$$

The intersection condition stipulates that:

$$N = K + a\mathbf{v} = P + b\mathbf{n}$$

which means that the point N is at the same time equal to the point K displaced by scaled vector $\mathbf{v}$ and to the point P displaced by scaled vector $\mathbf{n}$.

The intersection condition gives the following set of equations (one equation per dimension):

$$\begin{cases} k_x + a(l_x - k_x) = p_x + b(l_y - k_y) \\ k_y + a(l_y - k_y) = p_y + b(k_x - l_x) \end{cases}$$

which we can use to derive the formula for a by eliminating b and solving the resulting equation:

$$\begin{cases} k_x + a(l_x - k_x) - p_x = b(l_y - k_y) \\ k_y + a(l_y - k_y) - p_y = b(k_x - l_x) \end{cases}$$

$$\begin{cases} b = \dfrac{k_x + a(l_x - k_x) - p_x}{l_y - k_y} \\ b = \dfrac{k_y + a(l_y - k_y) - p_y}{k_x - l_x} \end{cases}$$

$$(k_x - l_x)(k_x + a(l_x - k_x) - p_x) = (l_y - k_y)(k_y + a(l_y - k_y) - p_y)$$

$$\begin{aligned} &(k_x - l_x)k_x + (k_x - l_x)a(l_x - k_x) - (k_x - l_x)p_x = \\ &\qquad (l_y - k_y)k_y + (l_y - k_y)a(l_y - k_y) - (l_y - k_y)p_y \end{aligned}$$

$$\begin{aligned} &(k_x - l_x)a(l_x - k_x) - (l_y - k_y)a(l_y - k_y) = \\ &\quad (l_y - k_y)k_y - (l_y - k_y)p_y - (k_x - l_x)k_x + (k_x - l_x)p_x \end{aligned}$$

$$
\begin{aligned}
&a((k_x - l_x)(l_x - k_x) - (l_y - k_y)(l_y - k_y)) = \\
&\quad (l_y - k_y)k_y - (l_y - k_y)p_y - (k_x - l_x)k_x + (k_x - l_x)p_x \\
a &= \frac{(l_y - k_y)k_y - (l_y - k_y)p_y - (k_x - l_x)k_x + (k_x - l_x)p_x}{(k_x - l_x)(l_x - k_x) - (l_y - k_y)(l_y - k_y)} \\
a &= \frac{(l_y - k_y)k_y - (l_y - k_y)p_y - (k_x - l_x)k_x + (k_x - l_x)p_x}{-(k_x - l_x)^2 - (l_y - k_y)^2} \\
a &= \frac{(k_x - l_x)(k_x - p_x) + (k_y - l_y)(k_y - p_y)}{(k_x - l_x)^2 + (l_y - k_y)^2} \qquad (3)
\end{aligned}
$$

The intersection point N can be now calculated as the point K translated by the appropriately scaled vector $\mathbf{v}$:

$$N = K + a\mathbf{v} \qquad (4)$$

This formula is used in the process of computing the normal distribution centered around the satellite path.

7. Nearest point on a polygonal chain

The piecewise-linear river Spree forms a *polygonal chain*, mathematically speaking. The nearest point on such a chain can be found by locating the nearest point on each line segment and taking the nearest of them all.

The nearest point on a particular line segment $\overline{KL}$ is found (almost) the same way as the nearest point on a line $\overleftrightarrow{KL}$. We use the formula (3) to obtain a. If $0 \leq a \leq 1$ then the point N from formula (4) is located within the line segment $\overline{KL}$ and is the point we look for. If $a < 0$ or $a > 1$ then the point N lies outside of the segment. In that case, the nearest point is either K or L. This is easy to determine.