Estimating KLD for log normal mixture model

August 1, 2018

In this paper I'll present a method to estimate KLD for mixture model of log-normal distribution, relevant for the brain research conducted by prof. Daphna Yoel.

1 The model

Our goal is to compute the KLD for both men and women populations. Our model assumes the following:

1. The total mean is 0
2. The total variance is 1
3. The population is composed of equal number of men and women

The null hypothesis (pure types) assumes the following conditional distributions:

$$f_{men} = \mathcal{N}(\xi, \sigma^2) \tag{1}$$
$$f_{women} = \mathcal{N}(-\xi, \sigma^2) \tag{2}$$

and $\sigma^2 = 1 - \xi^2$. The alternative hypothesis (mixture model) assumes the following conditional distributions:

$$f_{men} = p * \mathcal{N}(\xi + \varepsilon, \sigma^2) + (1-p) * \mathcal{N}(\xi - \delta, \sigma^2) \tag{3}$$
$$f_{women} = q * \mathcal{N}(\xi + \varepsilon, \sigma^2) + (1-q) * \mathcal{N}(\xi - \delta, \sigma^2) \tag{4}$$
$$\varepsilon, \delta > 0 \tag{5}$$

where p, q are the solution to the first assumption:

$$p * (\xi + \varepsilon) + (1 - p) * (\xi - \delta) = \xi$$
$$q * (\xi + \varepsilon) + (1 - q) * (\xi - \delta) = -\xi$$

solving the equation system yields:

$$p = \frac{\delta}{\varepsilon + \delta}$$
$$q = \frac{\delta - 2 * \xi}{\varepsilon + \delta}$$

with the additional constraint:

$$p, q > 0 \implies \delta > 2 * \xi \tag{6}$$

Next, to satisfy the second assumption we set:

$$\sigma^2 = 1 - (\xi - \delta)^2 - ((\xi + \varepsilon)^2 - (\xi - \delta)^2) * \frac{p + q}{2} \tag{7}$$

Since $\sigma^2 > 0 \implies (\xi - \delta)^2 - ((\xi + \varepsilon)^2 - (\xi - \delta)^2) * \frac{p+q}{2} < 1$ solving this equation yield another constraint:

$$1 > (\xi - \delta)^2 * (1 - \frac{p + q}{2}) + (\xi + \varepsilon)^2 * (\frac{p + q}{2}) \tag{8}$$

Setting p, q into the last equation yields:

$$(\xi - \delta)^2(2\varepsilon + \xi^2) + (\xi + \varepsilon)^2(2\delta - \xi^2) < 2(\varepsilon + \delta) \tag{9}$$

2 KLD computation

We estimate KLD using Monte Carlo simulation in the following manner. First we sample n observations from the mixture model distribution. Next we compute the likelihood under each scenario:

$$llk_{H_0} = log(\mathcal{N}(\bar{x}, s^2)) \tag{10}$$
$$llk_{H_1} = log(p_1 * \mathcal{N}(\xi - \delta, \sigma^2) + (1 - p_1) * \mathcal{N}(\xi - \varepsilon, \sigma^2)) \tag{11}$$

To compute KLD we use numeric integration:

$$KLD = \int_{-\infty}^{\infty} log(\frac{f_1(x)}{f_0(x)}) f_1(x) dx \tag{12}$$

We use the KLD to estimate the set of parameters $\{\xi, \delta, \varepsilon\}$ that a llrt will yield a 50% power.

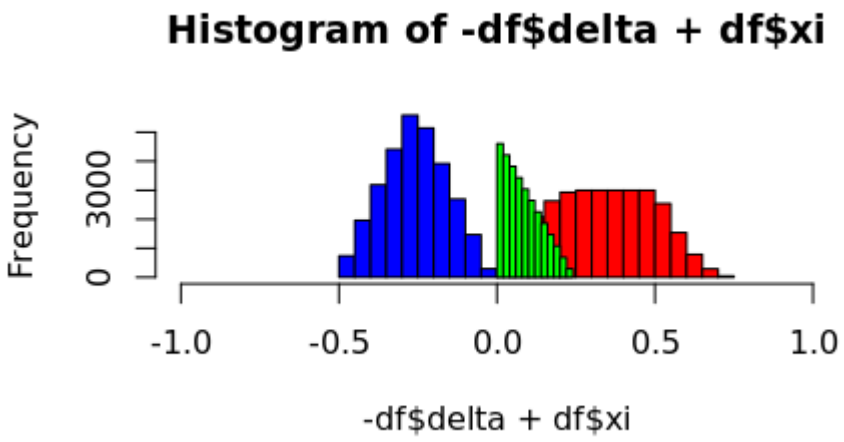
3 Results

In general the results of the computation draw a picture similar top the one we saw on the first simulation. A ridge like manifold of the KLD over the parameter set. The manifold follows an exponential like curve that peaks at the end of the feasible set at 0.2.

As we can see, this implies that a sample size of 20K is needed to make a discovery in this region.

3.1 Feasible parameter set

One thing worth mentioning is that the results reported above are partially due to the feasible set constraints. The following plot present the feasible set.



we can see that the feasible set define a family of mixture models that can be classified into families.

4 Conclusions and Discussion

The first conclusion we draw is that the mixture model is hard to detect when the centers are relatively close. We can only make a discovery when the centers are far away. This strengthen the motivation to use the non standardize data and use log-normal distribution.