Notes on Bayesian Information Criterion Calculation

July 14, 2012

Maximum likelihood estimate for the variance:

$$\sum_i \left(x_i - \mu_{(i)}\right)^2 = \sum_n \sum_{i \in D_n} \left(x_i - \mu(i)\right)^2 \tag{1}$$

The unbiased estimator for the variance is

$$\hat{\sigma}_n^2 = \frac{1}{R_n - 1} \sum_{i \in D_n} \left(x_i - \mu_n\right)^2 \tag{2}$$

Substitution (2) into (1) yields

$$\sum_i \left(x_i - \mu_{(i)}\right)^2 = \sum_n \left(R_n - 1\right) \hat{\sigma}_j^2 \tag{3}$$

Assuming the "identical spherical assumption" means

$$\hat{\sigma}_j^2 = \hat{\sigma}^2 \tag{4}$$

Then (3) becomes

$$\begin{aligned} \sum_i \left(x_i - \mu_{(i)}\right)^2 &= \left(\sum_n \left(R_n\right) - \sum_n (1)\right) \hat{\sigma}^2 \\ &= \left(R - K\right) \hat{\sigma}^2 \end{aligned} \tag{5}$$

Or, as written in the paper

$$\hat{\sigma}^2 = \frac{1}{R - K} \sum_i \left(x_i - \mu_{(i)}\right)^2 \tag{6}$$

The next step is to figure out the point probabilities under the maximum likelihood estimate $\hat{P}(x)$. Assuming the clusters are spherical gaussians, the probability for the position x_i in a cluster is

$$P(x_i) \propto \exp\left(-\frac{1}{2\sigma^2} \|x_i - \mu_{(i)}\|^2\right) \tag{7}$$

The constant of proportionality can be determined by computing the integral

$$\int P(x) \propto \int \exp\left(-\frac{1}{2\sigma^2}\|x-\mu\|^2\right) dx \tag{8}$$

$$= \int \exp\left(-\frac{1}{2\sigma^2}\sum_{\alpha=1}^{M}(x_\alpha-\mu_\alpha)^2\right)\prod_{\alpha=1}^{M} dx_\alpha \tag{9}$$

$$= \int \prod_{\alpha=1}^{M} \exp\left(-\frac{1}{2\sigma^2}(x_\alpha-\mu_\alpha)^2\right) dx_\alpha \tag{10}$$

$$= \prod_{\alpha=1}^{M} \int \exp\left(-\frac{1}{2\sigma^2}(x_\alpha-\mu_\alpha)^2\right) dx_\alpha \tag{11}$$

$$= \prod_{\alpha=1}^{M} \sqrt{2\pi\sigma^2} \tag{12}$$

$$= \left(2\pi\sigma^2\right)^{M/2} \tag{13}$$

So

$$P(x_i) = \sum_{n=1}^{K} \underbrace{P(x_i \in D_n)}_{\text{prob x is an element of cluster } D_n} \cdot P(x_i | x_i \in D_n) \tag{14}$$

The probability an element i is a member of cluster n, assuming identical cluster distributions is just the probability of picking an element of the cluster size R_n from the number of all possible points R.

$$P(x_i \in D_n) = \frac{R_n}{R} = \frac{R_{(i)}}{R} \tag{15}$$

The second factor is probability for a single cluster

$$P(x_i | x_i \in D_n) = \begin{cases} P_m(x_i) & \text{if } D_n = D_m \\ 0 & \text{if } D_n \neq D_m \end{cases}$$

The second term is the multivariate distribution as computed above

$$P(x_i | x_i \in D_n) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|x_i-\mu_n\|^2\right) \tag{16}$$

Combining (14), (15), and (16)

$$P(x_i) = \frac{R_n}{R}\frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|x_i-\mu_{(i)}\|^2\right)$$

$$= \frac{R_n}{R}\frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|x_i-\mu_{(i)}\|^2\right) \tag{17}$$

Converting to log-likelihoods. (Note that this, like all the logs in the paper, is log base-e. Different bases can be used, but they would lead to rescalings of some of the constants.)

$$
\begin{aligned}
l(D) &= \log \prod_i P(x_i) \\
&= \sum_i \log P(x_i) \\
&= \sum_i \left(\log \frac{R_n}{R} + \log \left(\frac{1}{(2\pi\sigma^2)^{M/2}} \right) - \frac{1}{2\sigma^2} \|x_i - \mu_{(i)}\|^2 \right) && (18) \\
&= \sum_{n=1}^{K} \sum_{x_i \in D_n} \left(\log \frac{R_n}{R} + \log \left(\frac{1}{(2\pi\sigma^2)^{M/2}} \right) - \frac{1}{2\sigma^2} \|x_i - \mu_{(i)}\|^2 \right) && (19) \\
&= \sum_{n=1}^{K} \left[R_n \left(\log \frac{R_n}{R} - \frac{M}{2} \log \left(2\pi\sigma^2\right) \right) - \frac{1}{2\sigma^2} \sum_{x_i \in D_n} \|x_i - \mu_{(i)}\|^2 \right] && (20)
\end{aligned}
$$

Now using the maximum likelihood assumption from (2),

$$
\begin{aligned}
\hat{l}(D) &= \sum_{n=1}^{K} \left[R_n \left(\log \frac{R_n}{R} - \frac{M}{2} \log \left(2\pi\hat{\sigma}^2\right) \right) - \frac{1}{2\hat{\sigma}^2} \left(R_n - 1\right) \hat{\sigma}^2 \right] && (21) \\
&= \sum_{n=1}^{K} \left[R_n \log R_n - R_n \log R - \frac{R_n M}{2} \log \left(2\pi\hat{\sigma}^2\right) - \frac{1}{2} \left(R_n - 1\right) \right] && (22)
\end{aligned}
$$

Using $\sum_{n=1}^{K} R_n = R$

$$
\hat{l}(D) = \sum_{n=1}^{K} R_n \log R_n - R \log R - \frac{RM}{2} \log \left(2\pi\hat{\sigma}^2\right) - \frac{1}{2} \left(R - K\right) \qquad (23)
$$

Now, consider two hypothesis, ϕ_1 and ϕ_2 (denoted M_j in the article, but I want to be clear this has nothing to do with the number of dimensions M). K, R_n, and σ are all functions of the models, ϕ. In our case, ϕ_1 is the clustering result after minimizing with a fixed number of clusters, and ϕ_2 is the result after splitting one of the clusters into two and doing k-means only over that original cluster. ϕ_2 is better than ϕ_1 if $BIC(\phi_2) > BIC(\phi_1)$.

For clarity, the maximum likelihood can be broken into the sum of two parts:

a model-dependent part and a model-independent part.

$$\hat{l}(D,\phi) = \sum_{n=1}^{K(\phi)} R_n(\phi)\log R_n(\phi) - R\log R - \frac{RM}{2}\log\left(2\pi\sigma(\hat{\phi})^2\right) - \frac{1}{2}\left(R - K(\phi)\right) \tag{24}$$

$$\begin{aligned} &= \left[\sum_{n=1}^{K(\phi)} R_n(\phi)\log R_n(\phi) - \frac{K(\phi)}{2} - \frac{RM}{2}\log\left(\sigma(\hat{\phi})^2\right)\right] \\ &\quad - \left[\frac{R}{2} + R\log R + \frac{RM}{2}\log 2\pi\right] \end{aligned} \tag{25}$$

$$= \hat{l}_{\text{model-dependent}}(D,\phi) + \hat{l}_{\text{model-independent}}(D) \tag{26}$$

Using the defintion of the BIC and eliminating the model-independent terms,

$$\begin{gathered} \hat{l}(D,\phi_2) - \frac{p_{\phi_2}}{2}\log R > \hat{l}(D,\phi_1) - \frac{p_{\phi_1}}{2}\log R \\ \hat{l}_{\text{model-dependent}}(D,\phi_2) - \frac{p_{\phi_2}}{2}\log R > \hat{l}_{\text{model-dependent}}(D,\phi_1) - \frac{p_{\phi_1}}{2}\log R \end{gathered} \tag{27}$$

This give a final test of the form

$$\begin{aligned} &\left[\sum_{n=1}^{K(\phi_2)} R_n(\phi_2)\log R_n(\phi_2) - \frac{K(\phi_2)}{2} - \frac{RM}{2}\log\left(\sigma(\hat{\phi_2})^2\right)\right] - \frac{p_{\phi_2}}{2}\log R \\ &> \left[\sum_{n=1}^{K(\phi_1)} R_n(\phi_1)\log R_n(\phi_1) - \frac{K(\phi_1)}{2} - \frac{RM}{2}\log\left(\sigma(\hat{\phi_1})^2\right)\right] - \frac{p_{\phi_1}}{2}\log R \end{aligned} \tag{28}$$

If this inequality holds, ϕ_2 is considered a better model the ϕ_1.