MA 324 Project
Misspecified Models

Group 1
Nallam Kalyan Sathvik 210123041
Gampa Sathvik 210123021
Pavan Adari 210123043

Aim

Aim of this project is to predict a distribution when we are given a sample of data. We will predict the distribution which is most likely the data is generated from. For this we will use three different criteria.

a) Kullback–Leibler Divergence
b) Akaike Information Criterion
c) Bayesian Information Criterion

Kullback–Leibler Divergence

Kullback–Leibler (KL) divergence (also called relative entropy), denoted $D_{KL}(P||Q)$, is a type of statistical distance: a measure of how one probability distribution P is different from a second, reference probability distribution Q. A simple interpretation of the KL divergence of P from Q is the expected excess surprise from using Q as a model when the actual distribution is P.

For discrete probability distributions P and Q defined on the same sample space,the relative entropy from Q to P is defined to be

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \, \log\left(\frac{P(x)}{Q(x)}\right),$$

For distributions P and Q of a continuous random variable, relative entropy is defined to be the integral

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \, \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x \,,$$

where p and q denote the probability densities of P and Q.

Akaike Information Criterion

The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model.

In estimating the amount of information lost by a model, AIC deals with the trade-off between the goodness of fit of the model and the simplicity of the model. In other words, AIC deals with both the risk of over-fitting and the risk of under-fitting.

Suppose that we have a statistical model of some data. Let *k* be the number of estimated parameters in the model. Let $\hat{L}$ be the maximized value of the likelihood function for the model. Then the AIC value of the model is the following

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Thus, AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages over-fitting, which is desired because increasing the number of parameters in the model almost always improves the goodness of the fit.

Bayesian Information Criterion

The Bayesian information criterion (BIC) or Schwarz information criterion (also SIC, SBC, SBIC) is a criterion for model selection among a finite set of models; models with lower BIC are generally preferred. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC).

- *When* fitting models, it is possible to increase the maximum likelihood by adding parameters, but doing so may result in over-fitting. Both BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC for sample sizes greater than 7.

The BIC is formally defined as

$$\mathrm{BIC} = k\ln(n) - 2\ln(\widehat{L}).$$

where

- $\hat{L}$ = the maximised value of the likelihood function of the model *M*,i.e.
  $\hat{L} = p(x \mid \hat{\theta}, M)$
- *n* = the number of data points in x.
- *k* = the number of parameters estimated by the model.

Methodology

We start with a set of candidate models $g_1, g_2, \ldots g_R$. The given data is approximated by the empirical PDF. Let it be denoted by *f*. Now we will calculate the information lost from using g_1 to represent *f* by calculating $D_{KL}(f||g_1)$. Similarly $D_{KL}(f||g_2)$….. $D_{KL}(f||g_R)$ are calculated. Now the candidate model with minimum KL divergence is chosen.

Similar methods are used for AIC and BIC as well where we chose the model which minimizes the value.

Observations

Four different sample distributions generated from normal, exponential, gamma and beta are taken.
And they are tested for those four distribution to predict the best model.
a) KL Divergence
For the exponential and normal samples beta distribution has the lowest KL divergence. So there is a misprediction in both these cases
b)AIC
AIC correctly predicted the model for all 4 samples.
c)BIC
BIC also correctly predicted the model for all 4 samples.
Note:See the python notebook for actual values.

Conclusion

Thus we can say that AIC and BIC are very good criteria for model selection but minimizing KL divergence may not always give the correct model.