Penalizing Likelihood with a BIC Score

David Gifford

March 20, 2017

When we perform model selection we will compare models that are appropriately penalized for their complexity and thus capacity. This is crucial because we can perfectly model every observation for a given data set with a suitably complex model. However, this "perfect" model would not generalize to other observations and thus would not capture the underlying biological mechanism we seek to understand. Thus we need a principled way to cause a model to "pay for" its complexity.

In our study of regression trees in Lecture 7 we needed to chose between two options at each tree leaf. Option 1 had us modeling the data under consideration with a single mean and variance. Option 2 has us modeling the data under consideration in two partitions each with a mean and variance, with the partitions being defined by a threshold on a factor value that was associated with each observation. Our Option 1 model has two parameters, a mean and variance. Our Option 2 model has five parameters: two means, two variances, and the factor value split point that defines the two partitions.

Our regression tree model assumes:

- Our entire data set X consists of n genes and m time points.
- The expression levels x_{it} of gene i at time t can be modeled by an normal distribution
- Each observation time point t is associated with p factor values f_{1t} through f_{pt}
- Model A: We can model all of the data under consideration S at a regression leaf with a single mean μ and variance σ^2

- Model B: Alternatively we can model the data under consideration at a regression tree leaf by first splitting it into two partitions S_1 and S_2 of times based on factor f_j where the members of S_1 have $f_{jt} \leq \tilde{f}_j$ and S_2 contains all of the other members.

Using the empirical maximum likelihood estimates for all means $\hat{\mu}$ and variances $\hat{\sigma}$ the likelihoods of Model A and Model B then are

$$\hat{\Theta}_1 = \{\hat{\mu}, \hat{\sigma}^2\} \tag{1}$$

$$\mathcal{L}_A(X|\hat{\Theta}_1) = \prod_{i=1}^{n} \prod_{t \in S} \mathcal{N}(x_{it}; \hat{\mu}, \hat{\sigma}^2) \tag{2}$$

$$\log \mathcal{L}_A(X|\hat{\Theta}_1) = n|S|\big[-\frac{1}{2} - \frac{1}{2}\log(2\pi\hat{\sigma}^2)\big] \tag{3}$$

$$\hat{\Theta}_2 = \{\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2, \tilde{f}_j\} \tag{4}$$

$$\mathcal{L}_B(X|\hat{\Theta}_2) = \prod_{i=1}^{n} \prod_{t \in S_1} \mathcal{N}(x_{it}; \hat{\mu}_1, \hat{\sigma}_1^2) \prod_{i=1}^{n} \prod_{t \in S_2} \mathcal{N}(x_{it}; \hat{\mu}_2, \hat{\sigma}_2^2) \tag{5}$$

$$\log \mathcal{L}_A(X|\hat{\Theta}_2) = n|S_1|\big[-\frac{1}{2} - \frac{1}{2}\log(2\pi\hat{\sigma_1}^2)\big] + n|S_2|\big[-\frac{1}{2} - \frac{1}{2}\log(2\pi\hat{\sigma_2}^2)\big] \tag{6}$$

Model B will always have a likelihood that is greater than or equal to Model A because Model B can fully represent Model A. Thus Model A is nested inside of the more general Model B.

Adopting a Bayesian viewpoint, a principled way to compare Model A and Model B on data set X would be to integrate the likelihood of each model over all of its possible parameters

$$\mathcal{L}_A(X) = \int P(X|\Theta_1)P(\Theta_1)d\Theta_1 \tag{7}$$

$$\mathcal{L}_B(X) = \int P(X|\Theta_2)P(\Theta_2)d\Theta_2 \tag{8}$$

While we do not have a general closed form solution to these integrals assuming uninformative $P(\Theta)$, we can approximate their value by adjusting

our our maximum likelihood estimate for a model with the Bayesian Information Criterion (BIC) to create an adjusted BIC score for a model. As the amount of data goes to infinity, the BIC score of the true underlying model will be provably higher than alternative models that are over/under-complex, assuming the true model is one of the models being scored. Thus, the BIC score permits an "apples to apples" comparison of two models that have different numbers of parameters. Unlike other methods such as χ^2 tests, Models A and B do not need to be nested when we use a BIC score to compare them. The BIC score is an approximation, and assumes that the true underlying distribution of X comes from an exponential family distribution. For full details, see Schwarz, Gideon E. (1978), "Estimating the dimension of a model", Annals of Statistics, 6 (2):461-464, doi:10.1214/aos/1176344136, MR 468014.

The general form of a BIC score for a model M when there are k parameters and n observations given the maximum likihood parameters $\hat{\Theta}$ for observations X

$$BIC_M = \log \mathcal{L}_M(X|\hat{\Theta}) - \frac{k}{2} \log n \tag{9}$$

Thus for our regression tree models, their respective BIC scores given their number of parameters (Model A: 2; Model B: 5) and the number of observations they describe (n genes at $|S|$ time points) are

$$BIC_A = \log \mathcal{L}_A(X|\hat{\Theta}_1) - \frac{2}{2} \log n|S| \tag{10}$$

$$BIC_B = \log \mathcal{L}_B(X|\hat{\Theta}_2) - \frac{5}{2} \log n|S| \tag{11}$$