A CONDITION FOR CONSISTENCY OF THE BIC - STAT 522B FINAL PROJECT

GAL AV-GAY - 48300057

1. SUMMARY

The Bayesian Information Criterion (Schwarz, 1978) is often used for variable selection, in that the preferred model in a model space is found by minimizing the BIC over the models in that space. Similar to the Akaike Information Criterion (AIC), the BIC penalizes complexity. The BIC has been criticized as being too liberal in that it selects unnecessary covariates when the covariate space is large. In their 2008 paper, "Extended Bayesian Information criterion for model selection with large model space", Chen and Chen introduce an extended family of Bayes Information Criteria that results in a more tightly controlled false-discovery rate than the ordinary BIC. The consistency of this extended BIC is established, allowing the number of covariates to approach infinity with the sample size, and in doing so, it is revealed that the ordinary BIC is likely inconsistent when the number of covariates $p_n > \sqrt{n}$, where n is the sample size.

In this article we attempt to prove that there is a positive probability of inconsistency when the number of covariates $p \geq \sqrt{n}$ for the ordinary BIC. We identify a counter-example where the true model, s_0, is empty. In this case a one-covariate model will be selected over the true model when any one of the one-covariate loglikelihoods exceeds the likelihood of the empty model by $0.5log(n)$. The goal is to show that the maximum inflation in the

Date: March 9th 2015.

log likelihood is of order $0.5log(n)$ when we set $P = \sqrt{n}$, meaning that a non-empty model will be selected with positive probability when $P \geq \sqrt{n}$.

In this counterexample we only get as far as proving a positive probability for inconsistency of the ordinary BIC at $P \geq \sqrt{nlog(n)}$. We take a guess that our distance from the true result comes from considering only one-covariate models as an alternative to the empty model, rather than all models in the space.

2. Bayesian Information Criterion

Let $\{(y_i, x_i) : i = 1, ..., n\}$ be independent observations. The conditional density of y_i given x_i is $f(y_i|x_i, \theta)$ where $\theta \in \Theta \subset \mathbb{R}^P$, with P being a positive integer. The likelihood function of θ is given by

$$L_n(\theta) = f(x; \theta) = \prod_{i=1}^{n} f(y_i|x_i, \theta)$$

Where $Y = (y_1, ..., y_n)$. Let s be a subset of $\{1, ..., P\}$. We denote by $\theta(s)$ the parameters θ such that those components outside s are set to zero or some known values. The BIC selects the model that minimizes the following quantity:

$$BIC(s) = -2logL_n\{\hat{\theta}(s)\} + v(s)log(n)$$

where $\hat{\theta}(s)$ is the maximum likelihood estimator of $\theta(s)$ and $v(s)$ is the number of components in s. The first term is a likelihood, while the second term is a penalty on the complexity of the model. Minimizing the BIC is a compromise between maximizing the likelihood and minimizing the number of covariates.

3. Proof

We begun by constructing a likelihood ratio test statistic. Under our null hypothesis, we have that the true model, s_0, is the empty one. Our

alternative hypothesis involves a particular one covariate model being the true model. We guess that it would be more informative to have a more composite alternative hypothesis. We are not considering all $\tilde{P}$-covariate models where $\tilde{P} \in \{2, ..., P\}$.

Our likelihood for our empty model is $L_n\{\phi\}$ and the likelihood under the one covariate model is $L_n\{s_1\}$. We write the BIC under each hypothesis as:

$$BIC(\phi) = -2logL_n\{\phi\}$$

$$BIC(s_1) = -2logL_n\{s_1\} + log(n)$$

Therefore we choose a particular one covariate model s_1 over ϕ when

$$logL_n\{s_1\} - logL_n\{\phi\} > \frac{1}{2}log(n)$$

We now construct our counterexample, with our null hypothesis being the empty model, and our alternative being a particular one covariate model such that,

$$H_0 : \ Y_i = \beta_0 + \epsilon_i$$

$$H_1 : \ Y_i = x_{1,i}\beta_1 + \beta_0 + \epsilon_i$$

Where $Y_i \sim i.i.d\ N(0, 1)$ and $\epsilon_i \sim N(0, 1)$. Under these two hypotheses, our log likelihoods are the following:

$$logL_n\{\phi\} = -\frac{1}{2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0)^2$$

$$logL_n\{s_1\} = -\frac{1}{2}\sum_{i=1}^{n}(Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_{1,i})^2$$

where

$$\hat{\beta}_0 = \bar{Y}$$

and so

$$logL_n\{\phi\} = -\frac{1}{2}(n-1)S_n^2$$

In addition,

$$\tilde{\beta}_0 = \bar{Y} - \tilde{\beta}_1\bar{X}, \quad \tilde{\beta}_1 = \frac{\sum_{i=1}^n (X_i - X)(Y_i)}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

And therefore we can write:

$$logL_n\{s_1\} = -\frac{1}{2}\sum_{i=1}^n (Y_i - \frac{\sum_{i=1}^n (X_i - X)(Y_i)}{\sum_{i=1}^n (X_i - \bar{X})^2}\bar{X} - \frac{\sum_{i=1}^n (X_i - X)(Y_i)}{\sum_{i=1}^n (X_i - \bar{X})^2}X_{1,i})^2$$

$$= -\frac{1}{2}\sum_{i=1}^n (Y_i - \frac{\sum_{i=1}^n (X_i - X)(Y_i)}{\sum_{i=1}^n (X_i - \bar{X})^2}(\bar{X} - X_{1,i}))^2$$

$$= -\frac{1}{2}\sum_{i=1}^n \left[(Y_i^2) - 2\frac{\sum_{i=1}^n (X_i - X)(Y_i)}{\sum_{i=1}^n (X_i - \bar{X})^2}(\bar{X} - X_{1,i})Y_i + \left(\frac{\sum_{i=1}^n (X_i - X)(Y_i)}{\sum_{i=1}^n (X_i - \bar{X})^2}(\bar{X} - X_{1,i})\right)^2\right]$$

$$= -\frac{1}{2}(n-1)S_n^2 - \frac{1}{2}\frac{(\sum_{i=1}^n (X_i - X)(Y_i))^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$= -\frac{1}{2}(n-1)S_n^2\left(1 - \frac{(\sum_{i=1}^n (X_i - X)(Y_i))^2}{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i)^2}\right)$$

$$= -\frac{1}{2}(n-1)S_n^2\left(1 - \rho_{XY}^2\right)$$

The number of one-covariate models under our alternative hypothesis is P. Therefore, the likelihood that at least one of the P one-covariate model likelihoods exceeds the likelihood of our true empty model s_0 by $0.5log(n)$ is equal to the probability that the maximum of P one-covariate likelihoods will exceed the likelihood of the empty model by $0.5log(n)$. We write this probability in the form below:

$$P\left\{max_{s=1,...P}\left(-\frac{1}{2}(n-1)S_n^2\left(1-\rho_{XY}^2\right)\right) - \left(-\frac{1}{2}(n-1)S_n^2\right) \geq 0.5log(n)\right\}$$

$$= P\left\{max_{s=1,...P}\left(-\frac{1}{2}(n-1)S_n^2\left(1-\rho_{XY}^2\right)\right) \geq 0.5log(n) - \left(\frac{1}{2}(n-1)S_n^2\right)\right\}$$

$$= 1 - P\left\{ max_{s=1,\dots P}\left(-\frac{1}{2}(n-1)S_n^2\left(1-\rho_{XY}^2\right)\right) < 0.5log(n) - \left(\frac{1}{2}(n-1)S_n^2\right)\right\}$$

Given that our observations are independent, we can write our probability as the following:

$$= 1 - P\left\{ -\frac{1}{2}(n-1)S_n^2\left(1-\rho_{XY}^2\right) < 0.5log(n) - \left(\frac{1}{2}(n+1)S_n^2\right)\right\}^P$$

$$= 1 - P\left\{ -\frac{1}{2}(n-1)S_n^2\left(1-\rho_{XY}^2\right) + \left(\frac{1}{2}(n+1)S_n^2\right) < 0.5log(n)\right\}^P$$

$$= 1 - P\left\{(n-1)S_n^2\rho_{XY}^2 < log(n)\right\}^P$$

$$= 1 - P\left\{(n-1)S_n^2\rho_{XY}^2 < log(n)\right\}^P$$

$$= 1 - P\left\{\frac{(\sum_{i=1}^n (X_i - X)(Y_i))^2}{\sum_{i=1}^n (X_i - \bar{X})^2} < \frac{log(n)}{n-1}\right\}^P$$

$$= 1 - P\left\{\tilde{\beta_1}^2 S_x^2 < \frac{log(n)}{n-1}\right\}^P$$

$$= 1 - P\left\{\left|\tilde{\beta_1} S_x\right| < \sqrt{\frac{log(n)}{n-1}}\right\}^P$$

At this point, for the purpose of this counter example, we regard S_x as a constant and set $S_x = 1$. Furthermore, since we will eventually be investigating the asymptotic result, we replace $n-1$ with n. Disclaimer: I am not so sure about how these steps can be justified concretely. We therefore have,

$$1 - P\left\{\sqrt{n}\left|\tilde{\beta_1}\right| < \sqrt{log(n)}\right\}^P$$

$$= 1 - P\left\{-\sqrt{log(n)} < \sqrt{n}\tilde{\beta_1} < \sqrt{log(n)}\right\}^P$$

$$= 1 - \left(P\left\{\sqrt{n}\tilde{\beta_1} < \sqrt{log(n)}\right\} - P\left\{\sqrt{n}\tilde{\beta_1} < -\sqrt{log(n)}\right\}\right)^P$$

We use the following result for the finite sample distribution of $\tilde{\beta_1}$, using the fact that the true value of $\tilde{\beta_1} = 0$, and that $\sigma^2 = 1$ for this example:

$$\sqrt{n}(\tilde{\beta_1} - 0) \sim N(0,1)$$

and so

$$1 - \left(\Phi(\sqrt{log(n)}) - \Phi(-\sqrt{log(n)})\right)^P$$
$$= 1 - \left(\Phi(\sqrt{log(n)}) - 1 + \Phi(\sqrt{log(n)})\right)^P$$
$$= 1 - \left(2\Phi(\sqrt{log(n)}) - 1\right)^P$$

Now, we approximate using the tail bounds of the normal distribution, for example

$$1 - \phi(x) = \frac{\phi(x)}{x}$$

I am not so sure about the assumptions and conditions for this approximation, but in any case we set $x = \sqrt{log(n)}$, so that we can write our probability as the following:

$$1 - \left(2\Phi(\sqrt{log(n)}) - 1\right)^P = 1 - \left(2(\Phi(\sqrt{log(n)}) - 1) + 1\right)^P$$

$$= 1 - \left(-2(1 - \Phi(\sqrt{log(n)})) + 1\right)^P = 1 - \left(-2(\frac{\Phi(\sqrt{log(n)})}{\sqrt{log(n)}}) + 1\right)^P$$

$$= 1 - \left(-2(\frac{e^{-\frac{1}{2}log(n)}}{\sqrt{log(n)}}) + 1\right)^P$$

$$= 1 - \left(1 - \frac{2}{\sqrt{nlog(n)}}\right)^P$$

We want to find the asymptotic behaviour of this probability, so we take the limit as $n \to \infty$

$$1 - \lim_{n\to\infty} \left(1 - \frac{2}{\sqrt{nlog(n)}}\right)^P$$

Now we use the following property of the exponential function,

$$e^{-x} = \lim_{n\to\infty} \left(1 - \frac{x}{n}\right)^n$$

we see that setting $P = \sqrt{nlog(n)}$, we obtain a positive probability

$$1 - \lim_{n\to\infty} \left(1 - \frac{2}{\sqrt{nlog(n)}}\right)^{\sqrt{nlog(n)}} = 1 - e^{-2} > 0$$

Where if we had set $P = \sqrt{x}$, this probability would have been zero.

4. CONCLUSION

We have shown that when $P \geq \sqrt{nlog(n)}$, a one covariate model will be chosen over the true, empty model, with positive probability. We guess that had we formulated a more composite alternative hypothesis, our result would have been stronger, as was mentioned in the paper.

5. BIBLIOGRAPHY

1. Chen, J. and Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model space. Biometrika, 95, 759-771.

2. Schwarz, G. (1978). Estimating the dimension of a model. Ann. Statist. 6, 461-4.