Minimum Description Length Model Selection Criteria for Generalized Linear Models

Mark H. Hansen and Bin Yu

Abstract

This paper derives several model selection criteria for generalized linear models (GLMs) following the principle of Minimum Description Length (MDL). We focus our attention on the mixture form of MDL. Normal or normal-inverse gamma distributions are used to construct the mixtures, depending on whether or not we choose to account for possible over-dispersion in the data. In the latter case, we apply Efron's [6] double exponential family characterization of GLMs. Standard Laplace approximations are then employed to derive computationally tractable selection rules. Each constructed criterion has adaptive penalties on model complexity, either explicitly or implicitly. Theoretical results for the normal linear model, and a set of simulations for logistic regression, illustrate that mixture MDL can "bridge" the selection "extremes" AIC and BIC in the sense that it can mimic the performance of either criterion, depending on which is best for the situation at hand.

Keywords: AIC; Bayesian methods; BIC; code length; information theory; minimum description length; model selection; generalized linear models

1 Introduction

Statistical model selection attempts to decide between competing model classes for a data set. As a principle, maximum likelihood is not well suited for this problem as it suggests choosing the largest model under consideration. Following this strategy, we tend to overfit the data and choose models that have poor predictive power. Model selection emerged as a field in the 1970s, introducing procedures that "corrected" the maximum likelihood approach. The most famous and widely used criteria are *An Information Criterion* (AIC) of Akaike [1, 2] and the *Bayesian Information Criterion* (BIC) of Schwarz [15]. They both take the form of a penalized maximized likelihood, but with different penalties: AIC adds 1 for each additional variable included in a model, while BIC adds $\log n/2$, where n is the sample size. Theoretical and simulation studies (*cf.* Shibata [16], Speed and Yu [18], and references therein), mostly in the regression case, have revealed that when the underlying model is finite-dimensional (specified by a finite number of parameters), BIC is preferred; but when it is infinite-dimensional, AIC performs best. Unfortunately, in practical applications we rarely have this level of