Parametric Empirical Bayes Model Selection - Some Theory, Methods and Simulation

Nitai Mukhopadhyay
Eli Lilly and Company

and

Jayanta Ghosh
Purdue University

Abstract

For nested models within the PEB framework of george and Foster (Biometrika,2000), we study the performance of AIC, BIC and several relatively new PEB rules under 0-1 and prediction loss, through asymptotics and simulation. By way of optimality we introduce a new notion of consistency for 0-1 loss and an oracle or lower bound for prediction loss.The BIC does badly , AIC does well for the prediction problem with least squares estimates. The structure and performance of PEB rules depend on the loss function. Properly chosen they rend to outperform other rules.

1 Introduction

Our starting point is a paper by George and Foster (2000), abbreviated henceforth as [6]. [6] propose a number of new methods using PEB (Parametric Empirical Bayes) ideas on model selection as a tool for selecting variables in a linear model. An attractive property of the new methods is that they use penalized likelihood rules with the penalty coefficient depending on data, unlike the classical AIC, due to Akaike (1973), and BIC, due to Schwartz (1978), which use constant penalty coefficients. The penalty for a model dimension q is usually λq, where λ is a penalty coefficient. [6] compare different methods through simulation.

Our major contribution is to supplement this with some theoretical work for both prediction loss and 0-1 loss. The former is supposed to be relevant in soft science, where one only wants to make good prediction, and the latter is relevant in hard science, where one wants to know the truth. It is known in model selection literature that these different goals lead to different notions of optimality.

Our theory is based on the assumption that we have nested , orthogonal models – a situation that would arise if one tries to fit an orthogonal polynomial of unknown degree. This special case receives special attention in [6].

Our paper is based on Chapter 4 of Mukhopadhyay (2000), subsequently referred to as [9]. A related paper is Berger, Ghosh and Mukhopadhyay, (2003), which shows the inadequacy of BIC in high dimensional problems.

The BIC was essentially developed as an approximation to the Bayesian integrated likelihood when all parameters in the likelihood have been integrated out. The model that maximizes this is the posterior mode, it minimizes the Bayes risk for 0-1 loss. It is shown in Berger, Ghosh and Mukhopadhyay, (2003) that BIC is a poor approximation to this in high dimensional problems.

The optimality of AIC in high dimensional prediction problems has been proved in a series of papers, e.g., Shibata (1981), Li (1987) and Shao (1997).

Both the BIC and AIC are often used in problems for which they were not developed.

We examine the penalties of [6] in Section 2 and make some alternative recommendations. All the model selection rules are studied in Sections 3 and 4 from the point of view of consistency under 0-1 loss.

In section 5 we follow the predictive approach, using the consistency results proved earlier. For the situation where least squares estimates are used for prediction after selection of a model, we define an oracle, a sort of lower bound, in the spirit of Shibata. In the PEB framework it is easy to calculate the limit of the oracle, namely, the function $B(\cdot)$ and show that the Bayes prediction rule and the AIC attain this lower bound asymptotically. This is not always the case for the PEB rules, which are Bayes rules for 0-1 loss.

Section 5 ends with a study of the case where Bayes (shrinkage) estimates are used instead of least squares estimates. Then the PEB rules are asymptotically optimal and can do substantially better than AIC. However, the benefit comes from the better estimates rather than more parsimonious model selection.

Simulations in Section 6, for both 0-1 and squared error prediction loss, bear out the validity of asymptotic results in finite samples, they also provide useful supplementary information.

Results similar to those outlined above are studied, in the Frequentist setting of Shao (1997), in Mukhopadhyay and Ghosh (2002) and for Shibata's Frequentist setting of nonparametric regression in Berger, Ghosh, and Mukhopadhyay, (2003). The assumptions, priors, results and proofs differ in the three cases. The PEB formulation of [6] provides a PEB background for the simplest as well as cleanest results of this type.

2 PEB Model Section Rules for 0-1 Loss

The problem of variable selection in nested orthogonal models can be put in the following canonical form in terms of the regression coefficients.

The data consist of independent r.v's $Y_{ij}, i = 1, 2, \cdots, p, j = 1, 2, \cdots, r$. There are p models $M_q, 1 \leq q \leq p$. Hardly any change occurs if $q = 0$ is also allowed. Under M_q,

$$Y_{ij} = \beta_i + \epsilon_{ij}, \quad 1 \leq i \leq q, \quad j = 1, 2, \cdots, r$$

$$= \epsilon_{ij}, \quad q+1 \leq i \leq p, \quad j = 1, 2, \cdots r,$$

with ϵ_{ij}'s i.i.d. $N(0, \sigma^2)$. For simplicity we assume σ^2 is known. If σ^2 is unknown the same theory applies if σ^2 is replaced by a consistent estimate of σ^2. If $r > 1$ and p is large, then a consistent estimate of σ^2 is available from the residuals $Y_{ij} - \bar{Y}_i$. In our asymptotics r is held fixed and $p \to \infty$. The sample size is $n = pr$. Clearly, the model M_q of dimension q specifies that $\beta_{q+1}, \cdots, \beta_p$ are all zero.

In the PEB formulation, see e.g. Morris (1983), the dimension of parameter space is reduced by assigning the parameters a prior distribution with a few unspecified (hyper-)parameters which are estimated from data and integrating out original parameters. [6] assume, as in Morris (1983), that $\beta_1, \cdots, \beta_q$ are i.i.d. $N(0, c\,\sigma^2/r)$. In our work we have used $c\,\sigma^2$, both choices have validity – see our discussion in Berger and Pericchi (2001). In any case in the simulations $r = 1$, so that our prior is the same as that of [6].

As indicated in Morris (1983), a PEB formulation is a compromise between a classical Frequentist approach and a full Bayesian approach.

In many decision theoretic examples based on real or simulated data, Efron and Morris (1973), Morris (1983) and others have shown that the PEB formulation permits borrowing of strength from estimates of similar parameters, leading to estimates that substantially improve classical estimates even in a Frequentist sense. However, this does not follow from PEB theory.

The PEB theory works well, i.e. provides better estimates than classical ones in the sense of cross-validation or being closer to a known true value, when the normality (or other prior) distributional assumption is checked by comparing the expected and empirical distribution of $\bar{Y}_i$'s. If M_q is true, then $\bar{Y}_1, \bar{Y}_2, \cdots, \bar{Y}_q$ are i.i.d. $N(0, c\sigma^2 + \sigma^2/r)$.

In the PEB formulation here there are two unknown remaining parameters, namely c and the true q denoted as q_0. The PEB solution adopted by us is to estimate c from data and put a prior $\pi(q)$ on q. We make one final assumption that $\sigma^2 = 1$ which can be ensured by a suitable scale transformation.

Suppose c is known and $\pi(q)$ is a prior on q. The Bayes solution is to maximize with respect to q. The likelihood with $\beta_1, \cdots, \beta_q$ integrated out namely,

$$L(q, c) = A\ \pi(q)(1 + rc)^{-q/2} \exp\{\frac{rc}{1 + rc} SSq\} \cdots \tag{1}$$

where $SSq = r\sum_{1}^{q} \bar{Y}_i^2$ and A doesn't depend on q or c. Since c is not known, one choice – referred to as a conditional maximum likelihood estimate of c – is to maximize the expression in (1) with respect to c, giving

$$r\hat{c}_q = max\{\frac{SS_q}{q} - 1, 0\} \tag{2}$$

We now take $\pi(q)$ uniform on $1 \leq q \leq p$. Then the PEB Bayes rule will choose M_q if q maximizes the expression in (1) after replacing c by $\hat{c}_q$. This amounts to maximizing with respect to q,

$$\Lambda(q) \equiv \Lambda(q, \hat{c}_q) = 2 \log L(q, c) = \frac{r\hat{c}_q}{1 + r\hat{c}_q} SSq - q \log(1 + r\hat{c}_q)$$

$$= SSq - q(1 + \log + \frac{SSq}{q}) \tag{3}$$

If instead of estimating c, we put a prior on c and then use Laplace approximation we should maximize

$$\Lambda^*(q) = \Lambda(q, \hat{c}_q) - \log q + 2 \log(\frac{SSq}{q}) \tag{4}$$

Details are given in [9].

Later we provide some evidence that a single estimate of a c across all models is preferable. A natural PEB estimate is obtained by taking $\pi(q) = 1/p$, and summing the expressions of the likelihood in (1) over $1 \leq q \leq p$ and then maximizing with respect to c. This estimate $\hat{c}_\pi$ is referred to as the Marginal Maximum Likelihood estimate in [6]. One then gets a third penalized (log) likelihood

$$\Lambda_\pi(q) = SSq - q\{\frac{1 + r\hat{c}_\pi}{r\hat{c}_\pi} \log_+(1 + r\hat{c}_\pi)\}$$

In this paper $\hat{c}_\pi$ will also stand for any estimate which converges a.s. to c as true $q_0 \to \infty$.

George and Foster [6] discuss the relative advantages and disadvantages of each estimate of c and refer to unpublished work of Johnstone and Silverman (2000).

The new model selection rules are to be compared with AIC which maximizes $SSq - 2q/r$ and BIC which maximizes $SSq - q\{\log(pr)\}/r$. As indicated before both these classical rules are inappropriate for high dimensional problems with 0-1 loss.

The rule based on $\Lambda(q)$ is essentially due to [6] except that, instead of our uniform prior, they choose the "binomial" prior.

$$\pi(q) = w^q(1 - w)^q \tag{4a}$$

where, according to [6], w is to be estimated also by maximizing (1). For a given q, it is clear that w appears only in the prior $\pi(q)$ and not on the likelihood of the data given M_q. The maximizing w, namely,

$$\hat{w}_q = q/p \tag{5}$$

can hardly be called a PEB estimate in the same spirit as $\hat{c}_q$. Also for q/p bounded away from zero an one, the penalty in (log) integrated likelihood due to this $\pi(q)$ is $0(q)$ whereas this part of the penalty vanishes at the end-points. In other words, irrespective of the data, the models in the middle range of q are being unduly penalized.

The binomial prior seems more appropriate in the all 2^p subsets model selection problem, where the models in the middle have cardinality $\binom{p}{q}$ which is much bigger than the cardinality of, say $q = 1$ or p.

Even for all subsets model selection, there is some confounding between w and c in the following sense. The Bayesian "non-centrality" parameter is

$$E(\sum_{1}^{p} \beta_i^2) = pwc \tag{6}$$

An estimate of this can only help determine the product wc. Separate estimation of w and c will require the use of the normal likelihood in a way that is not robust. We will return to this problem elsewhere.

3 Consistency

We first consider the case where c is known, so in the PEB criteria estimates $\hat{c}_q$, $\hat{c}_\pi$ are to be replaced by c.

It is clear that if M_{q_0} remains fixed (as $p \to \infty$), then the likelihood ratio of M_{q_0} with respect to any other fixed M_q, remains bounded away from zero and infinity. Hence it would be impossible to discriminate one of them from the other with error probabilities tending to zero as $p \to \infty$. That can happen only when $|q_0 - q_1| \to \infty$ as $p \to \infty$. The following definition is motivated by this fact.

Definition Let $q_0 \to \infty$ as $p \to \infty$. A penalized likelihood criterion $A(q, \underset{\sim}{Y}, p)$ for model selection is consistent at q_0 if given $\epsilon > 0$ and for sufficiently large p and q_0, there exists a k, (depending on ϵ, p, q_0, such that

$$P_{q_0}\{A(q_0, \underset{\sim}{Y}, p) > A(q, \underset{\sim}{Y}, p), \forall |q - q_0| \geq k\} > 1 - \epsilon \tag{7}$$

Of course we could take fixed q_0 and examine consistency from the right only. The treatment is exactly similar.

Let

$$A(q, \underset{\sim}{Y}, p) = SSq - q\lambda \tag{8}$$

for some $\lambda > 0$. Then for $q_1 > q_0$ and $q_1 - q_0 \to \infty$

$$A(q_0, \underset{\sim}{Y}, p) - A(q, \underset{\sim}{Y}, p) = -r \sum_{q_0+1}^{q} \bar{Y}_i^2 + (q_1 - q_0)\lambda = (q_1 - q_0)(\lambda - 1 + o_p) \tag{9}$$

Similarly for $q < q_1$ and $q_0 - q_1 \to \infty$,

$$A(q_0, \underset{\sim}{Y}, p) - A(q, \underset{\sim}{Y}, p) = (q_1 - q_0)(1 + rc - \lambda + o_p(1)) \tag{10}$$

We thus have

Proposition 3.1. The penalized likelihood criterion $A(q, \underset{\sim}{Y}, p)$ with constant penalty coefficient λ is consistent at all $q_0 \to \infty$ iff $1 < \lambda < 1 + rc$.

For AIC, $\lambda = 2$, so one would have consistency if $rc > 1$. If $rc < 1$, one can show that

$$A(1, \underset{\sim}{Y}, p) - A(q, \underset{\sim}{Y}, p) \to \infty \quad a.s. \tag{11}$$

if $q \to \infty$, i.e., AIC chooses M_1 or models not far from M_1. It is shown in section 5 that this is a good thing to do, if one wants to make predictions and least squares estimates are used.

The usual BIC with $\lambda = \log n$ is inconsistent, this extremely high penalty also leads to poor performance in prediction. A modified version due to several people, see [9] or Mukhopadhyay, Berger and Ghosh (2002) for references, has $\log p$ instead of $\log n$. That also is not consistent in general. For consistency one requires $r \geq 3$ and $1 + rc - \log r > 0$.

We now turn to the three PEB rules with estimates $\hat{c}_q$ or $\hat{c}_\pi$. It is easy to check that the rule based on $\Lambda_\pi(q)$ is consistent if $\hat{c}_\pi$ is a consistent estimate for c. To prove this we need to show

$$1 < \frac{1 + rc}{rc} log(1 + rc) < 1 + rc \tag{12}$$

The right hand inequality follows from

$$log(1 + rc) < rc \tag{13}$$

which is proved by the fact that the second derivative of $log(1 + x)$ is negative. The left hand inequality follows from

$$(1 + rc) log(1 + rc) > r \tag{14}$$

which is proved by the fact that the second derivative of $(1 + x) log(1 + x)$ is positive. The other two PEB criteria differ from each other by a quantity which is $o_p(q)$, hence they are either both consistent or both inconsistent. Since $\hat{c}_q$ has undesirable properties as an estimate of c (vide Section 4) neither of these rules is consistent in our sense. This does have some effect on their performance in prediction problems.

All one can show for these two cases is that $A(q_0, \underset{\sim}{Y}, p) - A(q, \underset{\sim}{Y}, p) \to \infty$ if $|q - q_0| \to \infty$ and (q_0/q_1) is bounded away from zero. To prove this, one has to use the behavior of $\hat{c}_q$ for $q > q_0$ which is studied in the next section.

4 Estimation of c.

By the law of large numbers, for large q,

$$\hat{c}_q = \frac{r\sum_1^q \bar{Y}_1^2}{q} - 1 = c \text{ (approximately), for } q \leq q_0$$
$$= \frac{q_0 c}{q} \text{(approximately)}, q > q_0 \tag{15}$$

Clearly, for large incorrect models, $\hat{c}_q$ decreases the penalty for each additional parameter, namely, 1+log $(1 + \hat{c}_q)$. This is counterintuitive. Plots of $\hat{c}_q$ for simulated data in [9] shows that $\hat{c}_q$ tends to die out for large incorrect values of q. This is the main reason why consistency became a problem for $\Lambda(q, \hat{c}_q)$.

If the true q_0 is fixed and not large, one cannot have a consistent estimate of c.

If $q_0 \to \infty$ at a rate faster than some known $\hat{q}$, then a consistent estimate is

$$\hat{c} = [r\sum_1^{\hat{q}} \bar{Y}_i^2 - 1]^+. \tag{16}$$

However such knowledge of $\hat{q}$ is unlikely. A plot of $\hat{c}_q$ provides good visual information about both c and true q_0.

An estimate of c, which is easy to calculate and has a nice Bayesian interpretation is the model average

$$\hat{c}_\pi = \sum_q \hat{\pi}_q \hat{c}_q \tag{17}$$

where

$$\hat{\pi}_q = \frac{e^{\Lambda(q,\hat{c}_q)}}{\sum_q e^{\Lambda(q,\hat{c}_q)}} \tag{18}$$

Asymptotic behavior of $\hat{c}_\pi$ is difficult to study. It is unlikely to be consistent in general for the following reason. For values of q much larger than q_0, $\hat{c}_q$ will be much smaller than c but such q's will have large weights $\hat{\pi}_q$ inappropriately. The net effect of this will be to pull down the average $\hat{c}_\pi$ away from c. Some evidence of this based on simulation is provided in [9].

We now make two rather strong assumptions which ensure consistency of a slightly modified version of $\hat{c}_\pi$.

A1) As $p \to \infty$, q_0/p is bounded away from zero

A2) There is a known positive number k such that $c \leq k$.

The modified version, also denoted by the same symbol, is

$$\hat{c}_\pi = \sum_1^p \hat{\pi}_q \min(\hat{c}_q, k) \tag{19}$$

Under our assumptions $\hat{c}_\pi \to c$ a.s. We sketch a proof. For slight simplicity, we take $r = 1$. For $q \leq q_0$

$$\hat{c}_q = c + O_p(q^{-\frac{1}{2}}) \tag{20}$$

This can be used to show for all $q < q_0(1-\epsilon), 0 < \epsilon < 1$, $\delta > 0$ and sufficiently small,

$$\Lambda(q, \hat{c}_q) - \Lambda(q_0, \hat{c}_{q_0}) < (q_0 - q_1\{\log(1+\hat{c}_{q_0}) - c + \delta\} + q_1\{\log(1+\hat{c}_{q_0}) - \log(1+\hat{c}_q)\} \tag{4.1}$$

$$< -(q_0 - q_1)\gamma, \tag{21}$$

(where $\gamma > 0$) with probability $> 1-\epsilon$. We have used the fact that $\log(1+c) < c$. We can now show as in the proof of Proposition 5.1 that

$$\sum_{q \leq q_0(1-\epsilon)} \exp\{\Lambda(q, \hat{c}_q) - \Lambda(q_0, \hat{c}_{q_0})\} \to 0 \tag{22}$$

with probability tending to one as $p \to \infty$. For $q \geq q_0$

$$1 + \hat{c}_q = \frac{q_0(1+\hat{c}_q)}{q_1} + (q_1 - q_0)(1 + r_q) \tag{23}$$

where by the strong law, $\sup_{q \geq q_0} |r_q| \to 0$ in probability. So, by concavity of $\log(x)$, there exists $\delta > 0$, such that for $p \geq q \geq q_0(1+\epsilon)$

$$\log(1+\hat{c}_q) \geq \frac{q_0}{q}\log(1+\hat{c}_{q_0}) + \delta + r_q \tag{24}$$

where r_q is a generic term such that $\sup_q |r_q|$ is $o_p(1)$. Then for $p \geq q > q_0(1+\epsilon)$,

$$\Lambda(q, \hat{c}_q) - \Lambda(q_0, \hat{c}_{q_0}) = (q-q_0)(1+r_q) + q_0(1+\log(1+\hat{c}_{q_0})) - q(1+\log(1+\hat{c}_q) \tag{25}$$

where $\sup_{q > q_0(1+\epsilon)} q_0^{-1}|r_q| = o_p(1)$

The expression in (25) is, by (24),

$$< -q\frac{\delta}{2}$$

Once again an analogue of (22) for $q > q_0(1+\epsilon)$ is true. So the contribution to $\hat{c}_\pi$ from $q > q_0(1+\epsilon)$ and $q < q_0(1-\epsilon)$ is negligible. But for $|q - q_0| < \epsilon$, $\hat{c}_q$ can be made as close to c by choice of ϵ. This proves the consistency of $\hat{c}_\pi$.

5 Bayes Rule for Prediction Loss and Asymptotic Performance

It is well-known (see, e.g., Shao (1997)) that the loss in predicting unobserved Y's, for an exact replicate of the given design, on the basis of given data is the sum of a term not depending on the model and the squared error loss $\sum_1^q(\bar{Y}_i-\beta_i)^2$. So in evaluating performance of a model selection rule it is customary to ignore the term not involving the model and focus on the squared error loss. We do so below.

For a fixed c the Bayes rule is described in the following theorem. We need to first define a quantile model. A model M_q is a posterior α-quantile model if $\pi(i+1 \leq q|\underset{\sim}{Y}) \leq \alpha < \pi(i \leq q|\underset{\sim}{Y})$ or equivalently.

Theorem 5.1. *The Bayes rule selects the smallest dimensional model if* $rc \leq 1$ *and the posterior* $\frac{rc-1}{2rc}$ *quantile model if* $rc > 1$

Proof Let M_q stand for the true (random) model with prior $\pi(q)$ The posterior distribution of β_i given M_q is

$$\pi(\beta_i|q,\underset{\sim}{Y}) = N(\frac{rc}{1+rc}\bar{Y}_i, c/(1+rc)), \quad i \leq q$$

$$= \text{point mass at zero}, \quad i > q$$

Hence

$$E\{(\bar{Y}_i-\beta_i)^2|q,\underset{\sim}{Y}\} = \{\frac{\bar{Y}}{1+rc}\}^2 + \frac{c}{1+rc}, \quad i \leq q$$

$$= \bar{Y}_i^2, \quad i > q$$

Similarly, $E\{(\beta_i-0)^2|q,\underset{\sim}{Y}\} = \{\frac{rc\bar{Y}_i}{1+rc}\}^2 + \frac{c}{1+rc}, \quad i \leq q$

$$= 0, i > q$$

Suppose we ignore the fact that we have to select from among nested models (i.e., we have to include all $j < i$ if we include i in our model) and just try to decide whether to set β_i non zero or zero. The posterior risks of these two decisions are

$$\Psi(i \text{ included } |\underset{\sim}{Y}) = \frac{c}{1+rc}\pi(q \geq i|\underset{\sim}{Y}) + \bar{Y}^2\{(\frac{1}{1+rc})^2\pi(q \geq i|\underset{\sim}{Y}) + (1-\pi(q \geq i|\underset{\sim}{Y})\},$$

$$\Psi(i \text{ excluded } |\underset{\sim}{Y}) = \frac{c}{1+rc}\{\pi(q \geq i|\underset{\sim}{Y})\} + \bar{Y}_i^2\{(\frac{rc}{1+rc})^2\pi(q \geq i|\underset{\sim}{Y})\}.$$

Hence inclusion of i is preferred iff

$$(\frac{1}{1+rc})^2\pi(q \geq i|\underset{\sim}{Y}) + (1-\pi(q \geq i|\underset{\sim}{Y})) < (\frac{rc}{1+rc})^2\pi(q \geq i|\underset{\sim}{Y})$$

which implies

$$\frac{1+rc}{2rc} < \pi(i \leq q|\underset{\sim}{Y})$$

Suppose $rc > 1$. Then we choose all i such that $\pi(i \leq q|\underset{\sim}{Y}) > \frac{rc-1}{1+rc}$. Given the obvious monotonicity of $\pi(i \leq q|\underset{\sim}{Y})$, this means we choose the $\frac{rc-1}{1+rc}$ posterior quantile model. Clearly this is the Bayes rule. More formally if $d(q_1)$ is the decision to choose model M_q, corresponding posterior risk

$$\Psi(q_1|\underset{\sim}{Y}) = \sum_{i=1}^{q_1} \Psi(i \text{ included}|\underset{\sim}{Y}) + \sum_{i=q_1+1}^{b} \Psi(i \text{ excluded}|\underset{\sim}{Y}).$$

$$\geq \sum_{1}^{p} \text{Min}\{\Psi(i \text{ included}|\underset{\sim}{Y}), \Psi(i \text{ excluded}|\underset{\sim}{Y})\} = \Psi(\frac{rc-1}{2rc}\text{quantile model}|\underset{\sim}{Y})$$

Similarly if $rc \leq 1$, it is easy to see that the simplest model minimizes the posterior risk among all models. This completes the proof.

To define asymptotic Empirical Bayes optimality, we define an oracle, i.e., a lower bound to the performance of any selection rule.

Let M_{q_0} be the true (unknown) model and $d(q_1)$ the decision to select M_{q_1}. Given $\underset{\sim}{Y}$, the PEB risk of $d(q_1)$ under M_{q_0}, after division by q_0, is

$$A(q_1) = \frac{1}{q_0}[\sum_{i=1}^{q_1} E\{(\bar{Y}_i - \beta_i)^2|q_0, \underset{\sim}{Y}\} + \sum_{i=q_1+1}^{p} E\{\beta_1^2|q_0, \underset{\sim}{Y}\}]$$

$$= \frac{c}{1+rc} + \frac{q_1}{q_0(1+rc)^2}\frac{1}{q_1}\sum_{i=1}^{q_1}\bar{Y}_i^2 + \frac{q_0-q_1}{q_0}\frac{c^2r^2}{(1+rc)^2}\frac{1}{q_0-q_1}\sum_{1+q}^{q_0}\bar{Y}_i$$

for $q_1 \leq q_0$

$$= \frac{c}{1+rc} + \frac{1}{(1+rc)^2}\frac{1}{q_0}\sum_{i=1}^{q_0}\bar{Y}_i^2 + \frac{q_1-q_0}{q_0}\frac{1}{q_1-q_0}\sum_{i=q_0+1}^{q}\bar{Y}_i^2$$

for $q_1 > q_0$.

Using the strong law of large numbers we obtain a heuristic approximation to $A(q_1)$ namely

$$\beta(q_1) = \frac{c}{1+rc} + \frac{q}{q_0(1+rc)}\frac{1}{r} + \frac{q_0-q}{q_0}\frac{c^2r^2}{(1+rc)r}, \quad q \leq q_0$$

$$= \frac{c}{1+rc} + \frac{1}{(1+rc)}\frac{1}{r} + \frac{q-q_0}{q_0}\frac{1}{r}, q_0 < q$$

which reduces to

$$\frac{c}{1+rc} + \frac{c^2r^2}{r(1+rc)} + \frac{q}{q_0}\frac{(1-rc)}{r} \quad q \leq q_0$$

and

$$\frac{c}{1+rc} + \frac{1}{r(1+rc)} + \frac{q-q_0}{q_0}\frac{1}{r} \quad q > q_0$$

Clearly $q_0\beta(\cdot)$ is a non-random approximation to the posterior risk under M_{q_0}. Note that $\beta(\cdot)$ is minimum at q_0 if $rc > 1$ and at $q = 1$ if $rc = 1$, then $\beta(\cdot)$ does not depend on q.

Theorem 5.2. *Let $A(\cdot)$ and $\beta(\cdot)$ be defined as above. Then*

$$\lim_{q_0\to\infty} \sup_q |A(q) - \beta(q)| = 0 \quad \textit{and} \lim_{q_0\to\infty} \frac{\inf_q A(q)}{\inf_q \beta(q)} = 1.$$

Proof We consider the case $q \le q_0$. The other case follows similarly.

$$\begin{aligned} A(q) - \beta(q) &= \frac{q}{q_0(1+rc)}^2 \{\frac{1}{q}\sum_1^q \bar{Y}_i^2 - (1+rc)\} \\ &\quad + \frac{q_0-q}{q_0}\frac{c^2r^2}{(1+rc)^2}\{\frac{1}{q_0-q}\sum_{q+1}^{q_0} \bar{Y}_i^2 - (1+rc)\} \\ &= T_1(q) + T_2(q) \end{aligned}$$

We show that $\sup_{q<q_0} |T_1(q)| \to 0$ a.s. One can show the other part $\to 0$ in a similar way.

By SLLN, given $\epsilon > 0$, we choose a Λ such that for $q > \Lambda$, $|\sum_1^q \bar{Y}_i^2/q - (1+c)| < \epsilon$. Since $q_0 > q$ and $(1+c)^2 > 1$, $|T_1(q)| < \epsilon$ for $\Lambda < q \le q_1$. The remaining $|T_1(q_0)|$ for $q \le \Lambda$ can be made smaller than ϵ if we choose q sufficiently large.

By repeated application of this kind of elementary argument one proves the first part of the theorem.

The first part implies

$$\lim_{q_0\to\infty} |\inf_q A(q) - \inf_q \beta(q)| = 0$$

Since,

$$\begin{aligned} \inf_q \beta(q) &= c \quad \text{if } c < 1 \\ &= 1 \quad \text{if } c \ge 1 \end{aligned}$$

is positive, the second part of the theorem follows.

Theorem 5.3. *For known c, the optimal model M_{q_c} is asymptotically equivalent to the oracle q minimizing $A(q)$ in terms of posterior predictive loss, i.e.,*

$$\frac{\textit{posterior predictive loss of } M_{q_c} \textit{ under } q_0}{(q_0 \inf_q A(q))} \to 1 a.s.$$
$$as \quad q_0 \to \infty$$

To prove this we need the following result, which has some independent interest.

Proposition 5.1 Let q_0 be the true model,. As $q_0 \to \infty$, $\pi(|q_1 - q_0| > \delta | \underset{\sim}{Y}) \to 0$ for any $\delta \to \infty$ such that $\delta = o(q_0)$.

This is in the spirit of posterior consistency at q_0 except that δ is not fixed but goes to infinity at a relatively slow rate.

Proof of Theorem 5.4 Without loss of generality take $r = 1$. If $c < 1$, the model M_{q_c} always chooses the simplest model. Hence its posterior risk (under q_0) is $q_0 A(q_c)$. Since $\beta(q)$ is minimized at $q = q_c$ in this case, we are done.

For $c > 1, \inf_q \beta(q) = 1$.

Also by Prop 5.1, $\frac{q_c}{q_0}$ a.s. $\to 1$

We consider the cases where $q_c \leq q_0$ The other case is similar. The posterior risk of M_{q_c} for $q_c < q_0$ is

$$\frac{c}{1+c} + \frac{1}{(1+c)^2} \frac{1}{q_0} \sum_1^{q_c} \bar{Y}_i^2 + \frac{c^2}{(1+c)^2} \frac{1}{q_0} \sum_{q_c+1}^{p} \bar{Y}_i^2$$

which $\to 1$ a.s. since $\frac{q_c}{q_0} \to 1$ a.s.

Proof of Prop. 5.1. We take $r = 1$ as before and let $\lambda(c) = \frac{1+c}{c} \log(1+c)$. It has been proved before that $1 < \lambda(c) < 1 + c$. Using the strong law, given $\epsilon > 0$, there exists $k > 0$ such that for $q > q_0 + k$, with probability tending to one

$$|(\Lambda(q) - \Lambda(q_0))|(q - q_0) - (1 - \lambda(c))| < \epsilon$$

i.e. $\Lambda(q) - \Lambda(q_0) < -(q - q_0)\gamma$, for some $\gamma > 0$.

Hence

$$\pi(q > q_0 + k | \underset{\sim}{Y}) \leq \sum_{q > q_0 + k} t^{(q - q_0)} \text{ where } t = e^{-\gamma}$$
$$= t^k/(1-t) \to 0$$

One can similarly show $\pi(q < q_0 - k | \underset{\sim}{Y}) \to 0$, using $\lambda(c) < 1 + c$

Remark 5.1. Theorem 5.1. holds for unknown c if $\hat{c}$ is a consistent estimate and we use $q_{\hat{c}}$ of the Empirica Bayes model selection rules but replacing $\hat{c}_q$ by $\hat{c}$. The same result holds for AIC also, which is interesting since AIC does not need to estimate c consistently. We prove this below.

One simply notes that in Section 3 we prove that for $rc > 1$, AIC is consistent for q_0, if $q_0 \to \infty$. Also for $rc < 1$, AIC $(q)-$ AIC $(1) \to -\infty$, if $q \to \infty$. Using

these facts one shows, as in the proof of Theorem 5.4., AIC attains the same risk as the oracle.

So far we have been looking at several Bayesian model selection rules from the point of view of prediction or squared error loss in a situation where after selection of model least squares estimates are used. Results differ in a major way if least squares estimated are replaced by the Bayes estimates $E(\beta_i|q, \underset{\sim}{Y}) = \frac{rc}{1+rc}\bar{Y}_i$ if M_q is chosen and $i \leq q$. Since the proofs are similar we merely state the main facts.

For a known c, the Bayes rule becomes the posterior median rule. This is a special case of a general result of Barbieri and Berger (2000) but can also be derived like Theorem 5.

To define a Bayesian oracle, we redefine

$$A(q) = \frac{1}{q_0}[\sum_{i=1}^{q} E\{(\beta i - \frac{rc}{1+rc}\bar{Y}_i)^2|q_0, \underset{\sim}{Y}\} + \sum_{q+1}^{p}\{(\beta_i - 0^2|q_0, \underset{\sim}{Y}\}]$$

$$= \frac{q}{q_0}\frac{c}{1+rc} + \frac{(q_0-q)}{q_0}\{\frac{c}{1+rc} + \frac{rc}{1+rc} + \sum_{q+1}^{q_0}\bar{Y}_i\}^2(qi_0 - q_1) \text{ if } q_1 \leq q_0$$

and

$$= \frac{c}{1+rc} + \frac{(q-q_0)}{q_0}\frac{c}{1+rc} + \sum_{q_0}^{q}(\frac{1}{1+rc}\bar{Y}_i)^2/(q_1 - q) \text{ if } q_1 > q_0$$

The heuristic nonrandom approximation is

$$\beta(q) = \frac{q}{q_0}\frac{c}{1+rc} + \frac{q_0-q}{q_0}\{\frac{c}{1+rc} + \frac{r^2c^2}{1+rc}\} \quad q \leq q_0$$

and

$$= \frac{c}{1+rc} + \frac{(q_1-q_0)}{q_0}\{\frac{c}{1+rc} + \frac{1}{1+rc}\} \quad q > q_0$$

$\inf \beta(q_1) = \frac{c}{1+rc}$, attained at q_0, for all c.

The posterior median Bayes rule as well as the PEB model selection rules followed by Bayes estimation attains the risk of the Bayesian oracle, namely q minimizing $A(q)$, provided c is known or a consistent estimate of c is used.

The advantage of using the (shrinkage) Bayes estimates can be seen comparing the $\inf \beta(q)$ for the two cases, namely $\frac{c}{1+rc}$ for Bayes estimates and $\frac{1}{r}$ for least squares estimates. For all fully Bayes rules reduce the posterior risk per component in the model by $\frac{1+rc}{rc}$ which can be very large if both r and c are small.

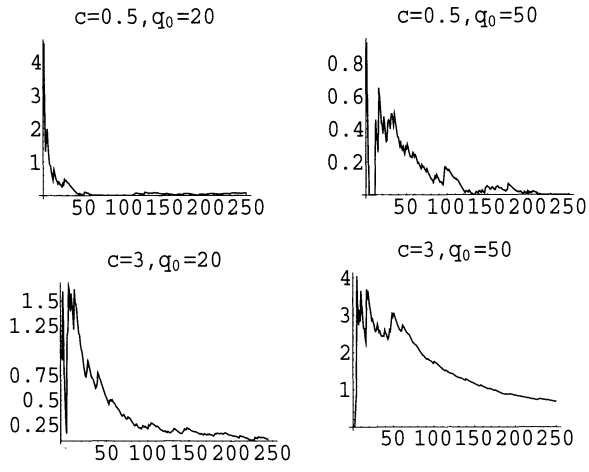

Figure 1: Behavior of $\hat{c}_q$ in a nested sequence of models.

6 Simulations and Discussion

A plot of c_q against q is a good Bayesian data analytic tool that provides information about both c and the true dimension q_0. This is true of all the four graphs in Figure 1 but it is specially noticeable when c if not too small.

The second set of simulations describe the performance of different model selection rules for 0-1 loss. We have taken $r = 1$ In addition to AIC, BIC and the three PEB rules defined in section 2, we consider the Conditional Maximum Likelihood rule (CML) of [6], in which both $\hat{c}_q$ and $\hat{w}_q$ are used as indicated in Section 2, even though the binomial prior seems unintuitive in the nested case.

In simulation $c = 0.5$ or 3. Higher values of c are considered in [9], the results are very similar to those for $c = 3$.

It is clear from Tables 1 and 2 that the BIC and CML are disastrous, as expected. AIC does well for $c = 3$ but badly for c =0.5, again as expected from Section 3. However, inconsistency is preferable to consistency in the prediction problem, vide the proof of Theorem 5.1 and Proposition 5.2. This is borne out by the third set of simulations.

The third set of simulations (Tables 3 and 4) describes performance of these criteria under prediction loss. Once again, $\Lambda^*(q)$ seems to do substantially better than $\Lambda(q)$ and Λ_π is somewhat worse than the other two. AIC is competitive for $c > 1$ and dramatically better than $c < 1$ This is because with least squares estimates neither of the three PEB rules are asymptotically optimal if $c < 1$. Of course the Bayes rule $q_{\hat{c}}$ for prediction loss would have done much better and be comparable to AIC.

q_0	4	5	10	20	40	500	800	900
$\Lambda(q)$	5 38 310	5 44 270	8 28 199	15 26 104	26 40 78	476 498 515	774 796 810	873 895 908
$\Lambda^*(q)$	1 3 7	1 3 10	2 4 10	2 8 20	3 21 40	475 497 510	771 795 809	873 895 908
$\Lambda_\pi(q)$	12 136 475	14 102 444	16 80 384	20 48 242	34 50 150	478 498 516	775 796 812	874 895 908
BIC	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
AIC	1 1 3	1 1 3	1 2 5	1 2 8	1 3 11	1 3 10	1 4 12	1 3 9
CML	1 1 999	1 1 999	1 1 999	1 1 999	1 999 999	999 999 999	999 999 999	999 999 999

Table 1: Quartiles of the dimensions selected by different criteria for $c = 0.5$, $r = 1$.

q_0	4	5	10	20	40	500	800	900
$\Lambda(q)$	3 5 22	4 5 12	8 10 11	17 20 21	37 39 41	497 500 500	797 799 800	897 899 900
$\Lambda^*(q)$	2 3 4	2 4 5	6 9 10	16 19 20	36 39 40	497 500 500	797 799 800	897 899 900
$\Lambda_\pi(q)$	4 8 64	4 5 38	8 10 14	17 20 21	37 39 41	497 500 500	797 799 800	897 899 900
BIC	1 1 2	1 1 2	1 1 3	1 1 3	1 1 3	1 1 3	1 1 3	1 1 3
AIC	2 4 4	3 4 5	6 9 10	16 19 20	36 39 40	497 499 500	797 799 800	896 899 900
CML	1 1 999	1 1 999	1 2 999	1 999 999	999 999 999	999 999 999	999 999 999	999 999 999

Table 2: Quartiles of the dimensions selected by different criteria for $c = 3$, $r = 1$.

q_0	4	5	10	20	40	500	800	900
$\Lambda(q)$	227.94	211.53	205.26	178.71	138.14	522.19	818.44	909.33
$\Lambda^*(q)$	35.77	20.66	37.06	42.47	54.76	518.25	816.34	908.1
$\Lambda_\pi(q)$	293.53	297.17	297.28	235.44	180.23	522.89	818.57	909.44
BIC	2.63	3.05	5.5	10.54	20.69	250.74	401.06	450.62
AIC	5.18	4.91	7.86	13.68	25.82	258.63	409.8	457.95
CML	425.15	412.92	466.09	499.04	574.25	998.05	1000.51	998.57

Table 3: Prediction loss of the models selected by different criteria for $c = 0.5$, $r = 1$.

q_0	4	5	10	20	40	500	800	900
$\Lambda(q)$	94.92	113.44	39.42	31.23	44.6	503.71	804.03	904.4
$\Lambda^*(q)$	14.36	19.09	15.6	24.45	44.22	503.65	804.08	904.36
$\Lambda_\pi(q)$	146.83	145.46	53.61	31.21	44.6	503.71	804.03	904.35
BIC	6.85	9.51	21.74	50.36	108.76	1489.84	2392.47	2693.92
AIC	6.56	7.09	13.13	23.23	43.62	503.66	804.32	904.23
CML	331.95	371.34	446.24	635.56	847.6	998.85	998.6	999.55

Table 4: Prediction loss of the models selected by different criteria for $c = 3$, $r = 1$.

We have not done any simulations on the posterior median Bayes rule, which uses PEB shrinkage Bayes estimates. It is expected to outperform AIC as seen from the comparison of $\beta(\cdot)'s$ for model selection followed by least squares and model selection followed by Bayes estimates. The three PEB criteria of Section 2, followed by Bayes estimates, are expected to do much better than evident in Tables 3 and 4 but not as well as the posterior median rule.

It may be worth pointing out that there is a basic difference between the median Bayes rule and AIC. Whether $c > 1$ or < 1, the median Bayes rule is consistent at q_0– a proof can be constructed using Proposition 5.1 But it then shrinks the estimates towards zero appropriately, depending on values of c. AIC doesn't have this option, it uses least squares estimates. So for critically small values of c, namely $c < 1$, it has to choose a much lower dimensional model to have some sort of shrinkage.

Bibliography

[1] Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petrov and F. Czaki, editors, Proceedings of the Second International Symposium on Information Theory, 267–271 Budapesk: Akad. Kiado.

[2] Barbieri, M. and Berger, J. (2000) Optimal Predictive Model Selection, ISDS Discussion Paper, Duke University.

[3] Berger, J.O., Ghosh J. K., and Mukhopadhyay, N. (2003) Approximations and consistency of Bayes factors as model dimension grows, Journal of Statistical Planning and inference, **[112]**, 241-258.

[4] Berger, J. O. and Pericchi, L. R. (2001) Objective Bayesian Methods for Model Selection: Introduction and Comparison, IMS Lecture Notes, (P. Lahiri editor) **38**, 135–203.

[5] Efron, B. and Morris, C. (1973) Stein's Estimation Rule and its Competitors an Empirical Bayes Approach, Journal of the American Statistical Association, **68**, 117-130.

[6] George, E. I. and Foster, D. F. (2000) Calibration and Empirical Bayes Variable Selection, Biometrika, **87**, 731–747.

[7] Li, K-C (1987) Asymptotic Optimality of c_p, c_l, cross Validation and Generalized cross Validation: Discrete Index Set, Annals of Statistics, **15**, 958-975.

[8] Morris, C. (1983) Parametric Empirical Bayes Inference, Journal of the American Statistical Association, **78**, 47-55.

[9] Mukhopadhyay, N. (2000) Bayesian Model Selection for High Dimensional Models with Prediction Loss and 0-1 loss, thesis submitted to Purdue University.

[10] Mukhopdhyay, N. and Ghosh, J.K. (2002) Bayes Rules for Prediction Loss and AIC, (submitted).

[11] Rissanen, J. (1983), A Universal Prior for Integers and Estimation by Minimum Description Length, Annals of Statistics, **11**, 416-431.

[12] Schwartz, G. (1978) Estimating the Dimension of a Model, The Annals of Statistics, **6**, 461-464.

[13] Shao, J. (1997) An Asymptotic Theory for Linear Model Selection, Statistica Sinica, **7**, 221-264.

[14] Shibata, R. (1981) An Optimal Selection of Regression Variables, Biometrika, **68**, 45-54.

[15] Shibata, R. (1983) Asymptotic Mean Efficiency of a Selection of Regression Variables, Annals of the Institute of Statistical Mathematics, **35**, 415-423.