Predictive Specification of Prior Model Probabilities in Variable Selection

By

Purushottam W. Laud

Medical ollege of i con in

and

Joseph G. Ibrahim

Harvard Univer ity

SUMM RY

We e amine the problem of specifying prior probabilities for all possible subset models in the conte t of variable selection in normal linear models. solution is proposed that uses a

where τ is a positive scalar parameter, and I is the $n \times n$ identity matri .

In selecting variables, we are interested in considering the 2^k possible models that can be obtained from (1.1) by retaining various subsets of the last k columns of the matri X, and modifying the length of β accordingly. To be specific, let m be a subset of the integers $\{0, \ldots, k\}$ containing 0, and let k_m denote the number of elements of m. Thus m identifies a model with an intercept and a specific choice of $k_m - 1$ predictor variables. With denoting the model space consisting of all 2^k models under consideration, we can write these as

$$Y = X_m \, \beta^{(m)} + \epsilon, \qquad m \in \quad , \tag{1.3}$$

where X_m is the $n \times k_m$ predictor matri under model m, and $\beta^{(m)}$ is the corresponding coefficient vector. Choosing one of the models in (1.3) is the goal of variable selection methods. The literature contains many techniques advanced for this purpose. See, for e ample, score prediction for each. Such predictions could, if appropriate, ta e guidance from some model, perhaps even outside , that was arrived at using past information. Similarly, a soil scientist may possess sufficient information and e pertise to ma e prior predictions on crop yield based on yields and covariates from the past, and a physician may be able to ma e individualized predictions of quantitative responses of patients in a study. In each case, it is desirable to incorporate the prior information and e pertise into the current analysis. To do this we require the investigator to ma e a prior prediction of the value of the response n-vector Y, ta ing into account all case-specific covariate information available. We denote this prediction by η, a fi ed vector regardless of the model under consideration. In eliciting priors, it has been recognized by many (Madigan, Gavrin and Raftery(1995) and the references there) that it is useful to focus attention on observable quantities as opposed to parameters. Such a focus becomes practically necessary in the case of model selection, where parameters abound.

Before proposing a prior distribution on , we briefly describe how L&I specify priors for $(\beta^{(m)}, \tau)$ for each $m \in$ by using η and a positive scalar c which quantifies the importance attached to the prior prediction η relative to the information in the data. Employing the normal-gamma conjugate family under each model, they ta e

$$\beta^{(m)}|\eta, \tau \sim No_{k_m}(\mu^{(m)}, \tau T_m) \ , \tag{2.1}$$

with

$$\mu^{(m)} = (X'_m X_m)^{-1} X'_m \eta \ , \tag{2.2}$$

$$Y|\tau, \eta \sim No_n(\eta, \gamma\tau I) \tag{2.5}$$

where $\gamma = c/(1+c)$. On the other hand, viewed through a model m and the prior (2.1) with (2.3),

$$Y|\tau, \eta \sim No_n(X_m\mu^{(m)}, \tau(I - (1-\gamma)P_m)) \tag{2.6}$$

where $P_m = X_m(X_m'X_m)^{-1}X_m'$ is the

$$p(m) = \frac{\left[\gamma_m \eta'(I-P_m)\eta + (\delta-2)^{-1}\lambda_m(n-k_m)\right]^{-n/2} e^{-k_m/2}}{\sum_{m\in\mathcal{M}} \left[\gamma_m \eta'(I-P_m)\eta + (\delta-2)^{-1}\lambda_m(n-k_m)\right]^{-n/2} e^{-k_m/2}} \; . \tag{2.10}$$

It is convenient here to ma e the choices

$$\lambda_m = l(n-k_m)^{-1}, \quad l > 0 \tag{2.11}$$

and

$$\gamma_m = b\alpha^{1/k_m}, \quad 0 \le b, \alpha \le 1 \; . \tag{2.12}$$

We observe that, with $\alpha = 0$ the prior probabilities for each fi ed k_m are equal. That is, we get uniform distributions over models of equal size. s $\alpha \to 1$, $p(m)$ can be dominated by $\eta'(I-P_m)\eta$ depending on b, δ and l. In practice, the e perimenter may choose $\eta \in C(X_{m^*})$ for some m^* due to the conte t of the e periment. Such a specification results in $\eta'(I-P_m)\eta = 0$ whenever $\eta \in C(X_m)$. This means relative probabilities for all models whose column spaces contain η depend only on λ and δ. Using the choices of δ and λ mentioned above, we have the following properties of the $p(m)$'s for such models : (i) ll models with the same number of predictors will get the same prior probability; (ii) For two models m and m', $k_{m'} > k_m$ implies $p(m') < p(m)$, thus giving larger probability to smaller models. We also note that with this choice of δ and λ, the prior mean and variance of τ both decrease as k_m increases. Thus larger models lead to smaller prior e pected precision. On both counts, these choices of δ and λ favor smaller models when their column spaces contain η.

If we ma e the choice $\alpha = 0$, it is clear from (2.10) that the prior probabilities are free of η and b. Moreover, by the definition of l following (2.10), they are also free of δ and l. Table 1 contains lists of these, a row for each choice of k up to 7. Each probability is followed, in parentheses, by the number of models over which it is spread evenly.

3 Examples

Before presenting two e amples to illustrate the priors of the previous section, we note that the specifications for η, δ, l, b and α can serve two purposes. Via (2.11) and (2.12), these generate a prior distribution on the model space . it the

Table 1: Prior Probabilities (Number of Models), $\alpha = 0$

	k_m							
k	1	2	3	4	5	6	7	8
1	0.622(1)	0.377(1)						
2	0.387(1)	0.470(2)	0.143(1)					
3	0.241(1)	0.438(3)	0.267(3)	0.054(1)				
4	0.150(1)	0.364(4)	0.330(6)	0.132(4)	0.020(1)			
5	0.093(1)	0.285(5)	0.340(10)	0.210(10)	0.065(5)	0.008(1)		
6	0.058(1)	0.210(6)	0.315(15)	0.260(20)	0.120(15)	0.030(6)	0.003(1)	
7	0.036(1)	0.154(7)	0.273(21)	0.280(35)	0.175(35)	0.063(21)	0.014(7)	0.001(1)

Together, a complete prior specification for the variable selection problem is achieved and, given the data y, one can compute posterior probabilities in a straightforward manner as

$$\begin{aligned} p(m|y) \quad \propto \quad & p(m) \times (n-k_m)^{-\delta/2}\, b^{k_m/2} \quad \times \\ & \Big[l(n-k_m)^{-1} + (y - P_m\eta)'(I - (1-\gamma_m)P_m)(y - P_m\eta) \Big]^{-\frac{n+\delta}{2}} . \end{aligned} \tag{3.1}$$

The choice $\alpha = 0$, $b = 1$ ma es this e pression free of the prior prediction η, reducing it to

$$p(m|y) \;\propto\; e^{-k_m/2}(n-k_m)^{-\delta/2} \Big[l(n-k_m)^{-1} + y'(I-P_m)y \Big]^{-\frac{n+\delta}{2}} . \tag{3.2}$$

Formally setting $l = \delta = 0$ now yields

$$p(m|y) \;\propto\; e^{-k_m/2} \left[y'(I - P_m)y \right]^{-n/2} . \tag{3.3}$$

This last e pression is just (2.8) written with the realized data y in place of the imaginary data Y_0. In other words, setting $\alpha = l = \delta = 0$ and $b = 1$ yields the posterior probabilities computed using the S&S priors for $(\beta^{(m)}, \tau)$ and a uniform distribution on . Such probabilities are, of course, in complete agreement with the local Bayes factors advanced in S&S.

Example 1 Wypij and Liu (1994) describe an e periment conducted to study personal e posure to ozone and how it relates to prevalent ozone concentrations and activities of individuals. Twenty three children were monitored for daytime e posure by means of a light-weight passive ozone sampler, newly developed by Koutra is et al.(1993). Each subject ept a diary of activities from 8 .M. to 8 P.M. Entries from these were aggregated and recorded on formatted sheets by field technicians. lthough the e periment involved other aspects such as validating measurements made by the new device, we describe here

Table 2: Model Probabilities,

of continous ozone concentration measurements made at an environmental data collection station within a reasonable distance (about 6 m) of the e perimental sites. Since the activity diaries contained hourly information, and the continuous measurements could be averaged correspondingly, it is possible to ma e a prior guess at the reponse variable values. In particular, let $X_6(k)$ denote the fraction of time spent indoors at home during the k^{th} hour. This could be determined from the individual diariesiTjﬂ76T91Tﬀlues.

Table 3: Model Probabilities, Hald Data with $\alpha = .602, b = .166$

Model	η_1	η_2	η_3	η_4
Intercept	.15 (.00)	.00 (.00)	.00 (.00)	.00 (.00)
x_2				

prior belief that the response variable does not have a regression relationship with any of the four predictors. These probabilities are also close to the noninformative specification obtainable from the row $k = 4$ of Table 1. Now it is nown from previous analyses appearing in the literature that the model with predictors X_1 and X_2 is quite adequate for these data. Table 3 reflects this in the model's substantially increased posterior probability in the η_1 column. lso, as we move to the column with prior prediction η_2 made with a belief in precisely this model, the prior probability attached to it has increased to 0.25. Moreover, the posterior probability is even higher. s we loo at the results under predictions η_3 and η_4, we see a decrease in the probabilities of this model, although it still remains more probable than any other. The prior probability of the model with X_1, X_4 shows an appreciable increase under η_3. However, the information in the data cause a shift away from this model, as reflected in the posterior.

Other calculations were carried out to see the behavior of these probabilities when the degree of belief in the prior predictions is increased. s e pected, there is an increase in the posterior probability of the model X_1, X_2 under the prior prediction η_2 as b and α increase. However, even under the e treme choice of unity for each, the posterior probability is 0.352. s b and α increase, the prior probability of this model increases to a ma imum of 0.342 and the ratio of posterior to prior probabilities decreases. Overall, the numerical e perience here see 4(s)-31000T7jfl118.0TDfle thatT7jfl1 elicitation predictive priors proposed in this article and in L&I show a desirable behavior a4(s)-31000he prior parameters are varied.

4 Discussion

Incorporating prior information into variable selection is not an easy tas . The available methods describe priors for the regression parameters in the various models under consideration, often concentrating on the noninformative case. See, for e ample, Mitchell and Beauchamp (1988) and the references therein. Here we have addressed the issue of specifying prior probabilities for the models. These are surmised from the prior prediction, η,T7jfl11ofonseT7jfl11 ariable alues in terpreted α, b, δ and l. The numerical results reported in Section 3 in0TDflethat the proposed priors could prove useful in practice.

In a recent paper, Madigan et al.(1995) demonstrate an elicitation of prior model probabilities in the conte t of graphical models by as ing an e pert to create imaginary cases with the aid of a randomizing program. This approach does not average over an imaginary replicate of the real e periment but uses elicited imaginary data in a Bayesian updating of uniform model probabilities. Yet, it is similar to this article in its focus on observable quantities. The article of Mitchell and Beauchamp (1988) contains an implicit specification of prior model probabilities in its equation (2.7). However, they recommend that the parameters of the prior be gleaned from the data. They also avoid computation of posterior probabilities, instead providing graphical summaries to assess the importance of various covariates.

The calculations of the posterior probabilities in Section 3 above employed the predictiv