Predictive Specification of Prior Model Probabilities in Variable Selection

By

Purushottam W. Laud

Medical College of Wisconsin

and

Joseph G. Ibrahim

Harvard University

SUMMARY

We examine the problem of specifying prior probabilities for all possible subset models in the context of variable selection in normal linear models. A solution is proposed that uses a prior prediction for the observable, an associated weight, and prior opinion regarding error precision as the only required input. Numerical examples are given to illustrate the method.

Keywords: Imaginary Experiment; Local Bayes Factors; Prior Prediction; Priors for Regression Coefficients.

1 Introduction

Variable selection in linear regression has drawn much attention in the literature. In this context the goal is to select a suitable subset from an available set of k predictors. To establish notation, consider the usual normal linear regression model

$$Y = X\beta + \epsilon \tag{1.1}$$

where Y is an n-vector of responses, X is the $n \times (k+1)$ full-rank matrix of fixed predictor variables with ith row $x_i' = (x_{i0}, x_{i1}, \ldots, x_{ik})$, $x_{i0} = 1$, $\beta = (\beta_0, \ldots, \beta_k)'$ is a $(k+1)$-vector of regression coefficients, and ϵ is an n-vector of random errors that is assumed to have a multivariate normal distribution with zero mean and precision matrix τI. Following the notation of Aitchison and Dunsmore (1975), we write

$$\epsilon|\tau \sim No_n(0, \tau I) \ , \tag{1.2}$$

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

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 matrix 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 $\mathcal{M}$ denoting the model space consisting of all 2^k models under consideration, we can write these as

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

where X_m is the $n \times k_m$ predictor matrix 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 example, Lindley (1968), Mallows (1973), Hocking (1976), and Lempers (1971). Additional references appear in the article by Mitchell and Beauchamp (1988). Recent articles on the topic include Laud and Ibrahim (1995) [henceforth referred to as L&I] and George and McCulloch (1993).

From the Bayesian viewpoint, the approach to the variable selection problem is, in principle, straightforward. The researcher needs to specify the prior probability of each model, a prior distribution for all of the parameters in each model, and compute the posterior probability of each model given the data. Such a prior must specify (i) a 2^k-long probability vector over $\mathcal{M}$, giving prior probability for each model and (ii) given any model m, a prior distribution for $(\beta^{(m)}, \tau)$. In this article we propose a new method to solve (i). We do this by focusing on observables, requiring only a few easily interpretable prior parameters to be specified. These same parameter specifications can also be used to solve (ii) as proposed in L&I.

2 Prior Distribution on the Model Space

In many practical situations, the investigator is able to focus better on observables rather than on parameters. For example, in studying performance on the Scholastic Aptitude Test (SAT) by high school students in a certain community, data from previous years may be available for the same community. With comparable covariate information at hand for each student that is planning to take the test this year, and also incorporating other specific knowledge about individual students, the investigator may be able to make a score prediction for each. Such predictions could, if appropriate, take guidance from some model, perhaps even outside $\mathcal{M}$, that was arrived at using past information. Similarly, a soil scientist may possess sufficient information and expertise to make prior predictions on crop yield based on yields and covariates from the past, and a physician may be able to make individualized predictions of quantitative responses of patients in a study. In each case, it is desirable to incorporate the prior information and expertise into the current analysis. To do this we require the investigator to make a prior prediction of the value of the response n-vector Y, taking into account all case-specific covariate information available. We denote this prediction by η, a fixed 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 $\mathcal{M}$, we briefly describe how L&I specify priors for $(\beta^{(m)}, \tau)$ for each $m \in \mathcal{M}$ 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 take

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

with

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

$$T_m = c \ X_m' X_m \tag{2.3}$$

and

$$\pi(\tau) \ \propto \ \tau^{\delta/2-1} \ exp\{-\lambda\tau/2\} \ , \tag{2.4}$$

where δ and λ are additional prior parameters. The motivation behind this specification, especially in the case of $\mu^{(m)}$, is that the prediction η should be respected as nearly as possible under each model m. On the one hand, if one were to write down a probability law for the observable Y free of all models in $\mathcal{M}$, one would consider (1.1) and (1.2) without the regression relation, thus replacing (1.1) by $Y \ = \ \mu \ + \epsilon$. Using the prior $\mu|\tau, \eta \sim No_n(\eta, c \ \tau I)$ yields

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

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 orthogonal projection matrix onto $C(X_m)$, the column space of X_m. The choice for $\mu^{(m)}$ given in (2.2) matches the first moments of the two distributions in (2.5) and (2.6) as closely as possible.

Turning to the main problem at hand, consider an imaginary past replicate of the current experiment having resulted in the response Y_0. Suppose that before this imaginary experiment, we started with a uniform distribution on $\mathcal{M}$, i.e., $p_0(m) = 2^{-k}$ for all $m \in \mathcal{M}$, and the prior on $(\beta^{(m)}, \tau)$ was taken to be the one implied in the derivation of local Bayes factors by Smith and Spiegelhalter (1980) [henceforth S&S], namely

$$\pi_0(\beta^{(m)}, \tau) \;\propto\; |X_m'X_m|^{1/2} e^{-k_m/2}(2\pi)^{-k_m/2}\tau^{k_m/2-1} \;. \tag{2.7}$$

Updating this prior using the imaginary data Y_0 yields

$$p(m|Y_0) = \frac{(Y_0'(I-P_m)Y_0)^{-n/2}\, e^{-k_m/2}}{\sum_{m\in\mathcal{M}} (Y_0'(I-P_m)Y_0)^{-n/2}\, e^{-k_m/2}} \;. \tag{2.8}$$

As Y_0 was not observed, these probabilities are to be viewed as random quantities that must be estimated or predicted. A natural choice is to average $p(m|Y_0)$ with respect to the distribution of Y_0 to obtain the desired probabilities $p(m)$. Now the distribution of Y_0 is just the right hand side of (2.5) since, in relation to the prediction η, the as-yet-unobserved entities Y_0 and Y can be considered exchangeable. Thus one should compute $p(m) = E[p(m|Y_0)]$ where $Y_0 \sim No_n(\eta, \gamma\tau I)$. However, this expectation does not have a closed analytic form. A convenient approximation can be obtained by replacing $Y_0'(I-P_m)Y_0$ by its expectation, leading to

$$p(m|\tau) \;\approx\; \frac{[\gamma\eta'(I-P_m)\eta + \tau^{-1}(n-k_m)]^{-n/2}\, e^{-k_m/2}}{\sum_{m\in\mathcal{M}} [\gamma\eta'(I-P_m)\eta + \tau^{-1}(n-k_m)]^{-n/2}\, e^{-k_m/2}} \;. \tag{2.9}$$

Finally, replacing τ here by its mode $\lambda^{-1}(\delta-2)$ under the prior (2.4) and allowing λ and γ to depend on m yield

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

It is convenient here to make 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 fixed k_m are equal. That is, we get uniform distributions over models of equal size. As $\alpha \to 1$, $p(m)$ can be dominated by $\eta'(I - P_m)\eta$ depending on b, δ and l. In practice, the experimenter may choose $\eta \in C(X_{m^*})$ for some m^* due to the context of the experiment. 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) All 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 expected precision. On both counts, these choices of δ and λ favor smaller models when their column spaces contain η.

If we make 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 examples 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 $\mathcal{M}$. Also, as in L&I, these assign prior distributions for the parameters of each model $m \in \mathcal{M}$ using equations (2.1) - (2.4).

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

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

The choice $\alpha = 0$, $b = 1$ makes this expression free of the prior prediction η, reducing it to

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

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 expression 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 $\mathcal{M}$. 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 experiment conducted to study personal exposure to ozone and how it relates to prevalent ozone concentrations and activities of individuals. Twenty three children were monitored for daytime exposure by means of a light-weight passive ozone sampler, newly developed by Koutrakis et al.(1993). Each subject kept a diary of activities from 8 A.M. to 8 P.M. Entries from these were aggregated and recorded on formatted sheets by field technicians. Although the experiment involved other aspects such as validating measurements made by the new device, we describe here

Table 2: Model Probabilities, Ozone Exposure Data

Model	Noninformative Prior Prior(Posterior) Probabilities	Informative Prior Prior(Posterior) Probabilities
X_1, X_4	.034 (.127)	.054 (.272)
X_1, X_4, X_5	.021 (.086)	.037 (.050)
X_1, X_2, X_4	.021 (.078)	.035 (.056)
X_1, X_3, X_4	.021 (.078)	.034 (.059)
X_1, X_2, X_5	.021 (.076)	.021 (.051)
X_1, X_5	.034 (.058)	.028 (.114)
X_1, X_3	.034 (.042)	.034 (.087)
X_4, X_5	.034 (.027)	.061 (.061)
X_1	.057 (.011)	.043 (.060)
Intercept only	.093 (.000)	.029 (.000)

only the parts that relate to model selection. The response variable, Y, was the 12-hour average personal ozone concentration (in parts per billion, ppb) for the subjects on different days. To build models for the prediction of this response, the authors considered the variables

indoor ozone concentration in the home of the subject, here denoted X_1,

outdoor concentration just outside the subject's home (X_2),

fraction of day spent outside the home (X_3).

Also included in the model search were the interaction terms $X_4 = X_2X_3$ and $X_5 = X_1(1 - X_3)$. There are 32 possible models, each including the intercept term.

Applying the techniques described in the previous section to the data from this experiment gave rise to the numbers in Table 2. The first two columns summarize the results of using the noninformative priors for model probabilities as well as for parameters within models, i.e., $\alpha = \delta = l = 0$, $b = 1$. The models listed include those receiving the top five posterior probabilities, the model with only the first covariate, and the model with no covariates. The model with the highest posterior probability is the one chosen by Wypij and Liu (1994) using various nonBayesian methods. By comparing posterior probabilities, it is clear that the top model is preferable to the one without any covariates. However, with other models also having similar posterior probabilities, the evidence for simply selecting the top model is less than convincing.

Now, in this experiment there was also some related information available in the form of continous ozone concentration measurements made at an environmental data collection station within a reasonable distance (about 6 km) of the experimental sites. Since the activity diaries contained hourly information, and the continuous measurements could be averaged correspondingly, it is possible to make 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 diaries. Also, the hourly values of the indoor and outdoor concentrations at each subject's home ($X_1(k)$ and $X_2(k)$) can be approximated by straightforward prorating schemes using the continuous measurements from the station and the 12-hour measurements from individual homes. We then use

$$\eta = (1/12) \sum_{k=1}^{12} \{X_1(k)X_6(k) + X_2(k)X_3(k)\}$$

as the guess at the response variable which directly measures the average 12-hour exposure for each individual. Wypij and Liu (1994) denote this η by X_2^H and give details for its calculation. They do not, however, treat it as a guess for the response. Instead, they use it in alternative models termed microenvironmental exposure models.

Having specified an informed prior guess at the response, we must now decide how much weight it should carry in relation to the actual response vector from the experiment. This weight γ_m is controlled by the parameters b and α via (2.12). Wishing, for the purposes of illustration, to keep γ_m between 0.10 and 0.15, we use the extremes in (2.12) to set $0.10 = b\alpha$ and $0.15 = b\alpha^{1/6}$. This implies $\alpha = 0.6417$ and $b = 0.1627$. Finally, the prior parameters for the precision must be specified. The instrument validation data reported in Wypij and Liu (1994) indicate a standard deviation of about 15ppb with a lower limit of about 10ppb. In terms of the precision parameter τ, we set $E[\tau] = \frac{1}{225}$ and $P(\tau \leq 0.01) = 0.95$. This results in $\delta = 5$ and $l = 70875$ (using $k_m = 1$), completing the prior specification.

The last column of Table 2 gives the prior and posterior probabilities computed using these choices and equations (2.10) and (3.1). Again, the listed models include those with the five highest posterior probabilities. In comparison with the noninformative case we see that the prior probabilities have changed somewhat, being larger for most of the listed models. Among these models, however, the prior probability is apportioned much in the same way in both columns. On the other hand, the posterior probabilities show a marked change. The top model now more clearly stands above its nearest competitors. Including related prior information in the analysis has resulted in a sharper distinction between models.

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_1	.09 (.00)	.00 (.00)	.00 (.00)	.00 (.00)
x_2	.09 (.00)	.00 (.00)	.00 (.00)	.00 (.00)
x_3	.09 (.00)	.00 (.00)	.00 (.00)	.00 (.00)
x_4	.09 (.00)	.00 (.00)	.00 (.00)	.00 (.00)
x_1, x_2	.06 (.29)	.25 (.34)	.20 (.25)	.23 (.30)
x_1, x_3	.06 (.00)	.00 (.00)	.00 (.00)	.00 (.00)
x_1, x_4	.06 (.17)	.07 (.01)	.14 (.03)	.11 (.02)
x_2, x_3	.06 (.00)	.00 (.00)	.00 (.00)	.01 (.00)
x_2, x_4	.06 (.00)	.00 (.00)	.00 (.00)	.00 (.00)
x_3, x_4	.06 (.01)	.06 (.00)	.03 (.00)	.07 (.00)
x_1, x_2, x_3	.03 (.14)	.15 (.22)	.15 (.24)	.15 (.24)
x_1, x_2, x_4	.03 (.14)	.15 (.23)	.15 (.25)	.14 (.23)
x_1, x_3, x_4	.03 (.13)	.14 (.15)	.14 (.17)	.12 (.14)
x_2, x_3, x_4	.03 (.07)	.09 (.01)	.09 (.01)	.08 (.01)
x_1, x_2, x_3, x_4	.02 (.07)	.09 (.05)	.09 (.06)	.09 (.05)

Example 2 Here we use the widely discussed Hald data (Hald, 1952 or Draper and Smith, 1981) mainly to see the effect of various choices of the prior prediction η and the weight given it via b and α. The data consist of four predictors measuring the percent composition of four ingredients of cement concrete, and a response variable measuring the heat evolved in calories per gram in thirteen samples. We consider four different prior predictions η_i, $i = 1, 2, 3, 4$. The first three are projections of the observed response vector on $C(1)$, $C(1, X_1, X_2)$, and $C(1, X_1, X_3)$, respectively, while the last is a perturbation of the observations, namely $\eta_4 = y + 2.445z$ where z has independent standard normal components and 2.445 is the root mean square error under the full model. The prior parameters for τ were taken to be $\delta = 25$ and $l = 1000$. Under the full model ($k_m = 5$), this amounts to approximately 95% prior probability that the precision is between 0.11 and 0.33 or that the variance is between 3.0 and 9.0. As in Example 1, we maintain γ_m between 0.10 and 015, here yielding $b = 0.166$ and $\alpha = 0.602$. Table 3 lists the prior probabilities (2.10) for each model with each of the four prior predictions. Corresponding posterior probabilities (3.1) are listed in parentheses.

There are many interesting features in this table. Because η_1 is in the column-space of the model having only the intercept term, this prediction is commensurate with the 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 known 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. Also, 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. As we look 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. As expected, 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 extreme choice of unity for each, the posterior probability is 0.352. As b and α increase, the prior probability of this model increases to a maximum of 0.342 and the ratio of posterior to prior probabilities decreases. Overall, the numerical experience here seems to indicate that the predictive specification of priors proposed in this article and in L&I show a desirable behavior as the prior parameters are varied.

4 Discussion

Incorporating prior information into variable selection is not an easy task. The available methods describe priors for the regression parameters in the various models under consideration, often concentrating on the noninformative case. See, for example, 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, of the response variable values along with the four easily interpreted scalars α, b, δ and l. The numerical results reported in Section 3 indicate that 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 context of graphical models by asking an expert to create imaginary cases with the aid of a randomizing program. This approach does not average over an imaginary replicate of the real experiment 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 predictive priors of L&I, thus using fully predictive priors. If other priors such as those in George and McCulloch (1993) are more convenient for the regression parameters, these can be used instead in conjunction with the prior model probabilities in (2.10).

The local Bayes factors of S&S are designed for model selection *without* incorporating any prior information. Equations (3.1)-(3.3) show that the S&S method can be recovered as the noninformative case of the fully predictive method proposed here. Although not related to this result, it is interesting to note that the priors inherent in using local Bayes factors do play a role in obtaining (2.10). They are used with the imaginary past data Y_0 and the resulting model probabilities are approximately averaged with respect to a prior predictive distribution.

References

[1] Aitchison, J., and Dunsmore, I. R. (1975), *Statistical Prediction Analysis*, New York : Cambridge University Press.

[2] Draper, N. R., and Smith, H. (1981), *Applied Regression Analysis*, (2nd ed.), New York : John Wiley.

[3] George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling", *Journal of the American Statistical Association*, 88, 881-889.

[4] Hald, A. (1952), *Statistical Theory With Engineering Applications*, New York : John Wiley.

[5] Hocking, R. R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1-51.

[6] Koutrakis, P., Wolfson, J. M., Bunyaviroch, A., Froelich, S. E., Kirano, K., and Mulik, J. D. (1993), "Measurement of Ambient Ozone using a Nitrite-coated Filter," *Analytical Chemistry*, 65, 209-214.

[7] Laud, P. W., and Ibrahim, J. G. (1995), "Predictive Model Selection", *Journal of the Royal Statistical Society*, Ser. B, 57, 247-262.

[8] Lempers, F. B. (1971), *Posterior Probabilities of Alternative Linear Models*, Rotterdam : Rotterdam University Press.

[9] Lindley, D. V. (1968), "The Choice of Variables in Multiple Regression" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 30, 31-66.

[10] Madigan, D., Gavrin, J., and Raftery, A. (1995), "Eliciting Prior Information to Enhance the Predictive Performance of Bayesian Graphical Models," *Communications in Statistics: Theory and Methods*, to appear.

[11] Mallows, C. L. (1973), "Some Comments on C_p, " *Technometrics*, 15, 661-675.

[12] Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," (with discussion), *Journal of the American Statistical Association* , 83, 1023-1036.

[13] Smith, A. F. M., and Spiegelhalter, D. J. (1980), "Bayes Factors and Choice Criteria for Linear Models," *Journal of the Royal Statistical Society*, Ser. B., 42, 213-220.

[14] Wypij, D., and Liu, L. J. S. (1994), "Prediction Models for Personal Ozone Exposure Assessment," pp 41-56 in *Case Studies in Biometry*, Lange et al. editors, John Wiley & Sons, NY.