Variable Selection and Model Comparison in Regresssion

John Geweke

University of Minnesota and
Federal Reserve Bank of Minneapolis

May 20, 1994

Abstract

In the specification of linear regression models it is common to indicate a list of candidate variables from which a subset enters the model with nonzero coefficients. In some cases any combination of variables may enter, but in others certain necessary conditions must be satisfied: *e.g.*, in time series applications it is common to allow a lagged variable only if all shorter lags for the same variable also enter. This paper interprets this specification as a mixed continuous-discrete prior distribution for coefficient values. It then utilizes a Gibbs sampler to construct posterior moments. It is shown how this method can incorporate sign constraints and provide posterior probabilities for all possible subsets of regressors. The methods are illustrated using some standard data sets.

Partial financial support from NSF grant SES-9210070 is gratefully acknowledged. Thanks to Rob McCulloch for providing some of the data used in this paper. This research was conducted while the author was a visitor at the Federal Reserve Bank of Minneapolis. The views expressed in this paper are those of the author and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

1. Introduction

The purpose of this paper is to propose and illustrate a new technique for an old and recurring problem, that of variable selection in linear regression. Loosely speaking, the task is to find the subsets of a prespecified set of potential covariates that best describe a dependent variable. Model selection and stepwise procedures address this problem; see Miller (1990) for a review and comprehensive bibliography of these procedures. This paper takes an explicitly subjective Bayesian view both of linear regression and the selection problem. The linear regression model is a predictive device -- its parameters are artificial, not real. As in Mitchell and Beauchamp (1986, 1988), prior distributions of parameters may be regarded as frequencies within a population of equally credible prediction experts. This explicitly includes the probability that a coefficient is zero, which is the proportion of experts who would omit the corresponding variable from the model.

This subjectivist interpretation carries with it no presumption of conjugacy in the priors. Just the opposite is true: prior information rarely establishes the link between coefficients and disturbance variance that is essential to methods exploiting conjugacy (Poirier, 1985). This paper proposes an independent prior distribution for each coefficient that is a mixture of a point mass at zero and a possibly truncated univariate normal distribution. These distributions are completely subjective, relying on no preprocessing of the data or other methods that destroy the independence of the prior information and the stochastic terms in the model. Through a series of examples, the paper illustrates that elicitation of prior distributions in this family is a natural procedure for a subjective Bayesian.

The work here is most closely related to George and McCulloch (1993). It is different from their work in four respects. First, the present paper employs subjective priors, whereas George and McCulloch employ a "semiautomatic approach" in which the prior incorporates sufficient statistics from the regression. Second, the prior distribution includes the possibility that variables are literally excluded from the model, whereas George and McCulloch for technical reasons utilize absolutely continuous prior cumulative distribution functions. Third, this paper avoids a computational shortcut utilized by George and McCulloch that entails assuming that certain coefficients are known *a priori* to be equal to their least squares estimates. (Both papers use the Gibbs sampling algorithm to carry out the computations. To solve the technical problems associated with a nonzero probability that coefficients are zero, a different version of the algorithm is employed here.) Finally, the

paper takes up the very common problem of ordered, or contingent, model selection which George and McCulloch do not address.

Prior and posterior distributions, and the computational algorithm, are outlined in the next section. Construction of prior distributions and several aspects of the posterior are illustrated in Section 3 through the same three examples used in George and McCulloch; this also affords some comparisons of the performance of the two procedures. Extension of these methods to contingent variable selection, which often arises in time series models, is developed in Section 4. Examples of this procedure (only one, in this draft) are provided in Section 5. The last section summarizes and discusses some possible extensions of this work.

2. Noncontingent variable selection

This section considers the standard regression variable selection problem, with proper, informative prior distributions for all parameters. In the standard problem, k^* out of k parameters each have a nonzero coefficient with prior probability 1, while there is positive probability that any combination of coefficients of the remaining $k-k^*$ variables have coefficients equal to zero. Thus if by "model" is meant a specific combination of coefficients whose posterior probability of being nonzero is positive, there are 2^{k-k^*} alternative models entertained by the prior distribution.

Here we treat in detail a simple but frequently arising instance of the standard selection problem. First, the regression model is linear in coefficients. Second, disturbances are normally distributed. Third, in the prior distribution all parameters are mutually independent, and for $k-k^*$ of the coefficients there is positive prior probability that the coefficient is zero; even more specifically, we develop the method for coefficient prior distributions that are mixtures of normal or truncated normal distributions, and discrete mass at the point 0. These assumptions may all be weakened; discussion of productive directions for weakening is deferred to the final section.

2.1 Prior and posterior distributions

In standard notation

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim \mathrm{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right) \tag{2.1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations on a dependent variable and $\mathbf{X}$ is an $n \times k$ matrix of n corresponding observations on k covariates. The likelihood function for this model is

$$\mathrm{L}(\beta,\sigma)=\sigma^{-n}\exp\left[-(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)/2\sigma^2\right]$$
$$=\sigma^{-n}\exp\left[-(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})/2\sigma^2\right]\exp\left[-(\beta-\mathbf{b})'\mathbf{X}'\mathbf{X}(\beta-\mathbf{b})/2\sigma^2\right] \quad (2.2)$$

where $\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ denotes the solution of the classic least squares problem. The k covariates include all of the regressors considered for inclusion in the model. Excluding a regressor means that the corresponding coefficient is zero in (2.1). It does not entail reducing the dimension of $\mathbf{X}$, which would render the distinction between the models meaningless as discussed by Poirier (1985, p. 712).

The investigator's prior distribution for each of the coefficients and the parameter σ are mutually independent. With prior probability $\underline{p}_i$, $\beta_i=0$; conditional on $\beta_i\neq 0$ the prior distribution of β_i is $\mathrm{N}(\underline{\beta}_i,\tau_i^2)$, possibly truncated to the interval (λ_i,υ_i):

$$d\Pi_i(\beta_i)=\underline{p}_i\,d\,\mathrm{H}_i(\beta)$$
$$+(1-\underline{p}_i)(2\pi)^{-1/2}\tau_i^{-1}\left[\Phi\left(\frac{\upsilon_i-\underline{\beta}_i}{\tau_i}\right)-\Phi\left(\frac{\lambda_i-\underline{\beta}_i}{\tau_i}\right)\right]^{-1}\exp\left[(\beta_i-\underline{\beta}_i)^2/2\tau_i^2\right]\mathrm{I}_{(\lambda_i,\upsilon_i)}(\beta_i) \quad (2.3)$$

where $\Pi_i(\cdot)$ denotes the prior c.d.f. of β_i; $\mathrm{H}(x)=0$ if $x<0$ and $\mathrm{H}(x)=1$ if $x\geq 0$; $\mathrm{I}_S(x)=1$ if $x\in S$ and $\mathrm{I}_S(x)=0$ if $x\notin S$; $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution; $0<\tau_i<\infty$; $-\infty\leq\lambda_i<\upsilon_i<\infty$; and $-\infty<\underline{\beta}_i<\infty$. The prior distribution of σ is of the standard form,

$$\underline{\nu}\underline{\sigma}^2/\sigma^2\sim\chi^2(\underline{\nu}). \quad (2.4)$$

The prior distribution is therefore proper and informative but nonconjugate. We choose this form because it is relatively easy to elicit one's subjective prior distribution about the coefficients in this form, yet the computational problem remains fairly simple. (Illustrations of prior construction are provided in Section 3.) The prior distribution is trivially coherent: *i.e.*, the prior distributions of nested models can be obtained as restrictions on each other.

The posterior distribution may be expressed up to a constant by combining (2.2), (2.3) and (2.4) in the usual way, but this expression is not particularly useful either for performing the computations or understanding the relation between the prior and posterior distributions. Instead we move directly to some more informative conditional distributions and the computational method based on them.

2.2 Computation

The computational procedure employed here is a Gibbs sampler with complete blocking. A value for each coefficient β_j is drawn in turn from its distribution conditional

on $\beta_l\,(l \neq j)$ and σ, and a value for σ is drawn conditional on β. In the algorithm the Gibbs sampler moves from any point in the support of β and σ to any nondegenerate neighborhood of any other point in the support with positive probability in one step. Convergence of the continuous state Markov chain induced by the Gibbs sampler to the posterior distribution may therefore be demonstrated following the argument of Tierney (1991).

The conditional distributions involved in the algorithm are simple. Given $\beta_l\,(l \neq j)$ and σ, define $z_i = y_i - \sum_{l \neq j} \beta_l x_{il}$. The conditional distribution of β_j follows from the simplified model

$$z_i = \beta_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \overset{IID}{\sim} \mathrm{N}(0, \sigma^2) \quad (i = 1, \ldots, n).$$

The likelihood function kernel is

$$\exp\left[-\sum_{i=1}^{n} \left(z_i - \beta_j x_{ij}\right)^2 \Big/ 2\sigma^2\right].$$

Conditional on $\beta_j = 0$ the value of the kernel is

$$\exp\left[-\sum_{i=1}^{n} z_i^2 \Big/ 2\sigma^2\right] \tag{2.5}$$

Conditional on $\beta_j \neq 0$ the corresponding kernel density for β_j is

$$\exp\left[-\sum_{i=1}^{n} \left(z_i - \beta_j x_{ij}\right)^2 \Big/ 2\sigma^2\right]$$
$$\cdot (2\pi)^{-1/2} \tau_j^{-1} \left[\Phi\left[\left(\upsilon_j - \underline{\beta}_j\right) \Big/ \tau_j\right] - \Phi\left[\left(\lambda_j - \underline{\beta}_j\right) \Big/ \tau_j\right]\right]^{-1} \exp\left[-\left(\beta_j - \underline{\beta}_j\right)^2 \Big/ 2\tau_j^2\right] \mathrm{I}_{(\lambda_j, \upsilon_j)}(\beta_j)$$

$$= \exp\left[-\sum_{i=1}^{n} \left(z_i - b x_{ij}\right)^2 \Big/ 2\sigma^2\right] \exp\left[-\left(\beta_j - b\right)^2 \Big/ 2\omega^2 - \left(\beta_j - \underline{\beta}_j\right)^2 \Big/ 2\tau_j^2\right]$$
$$\cdot (2\pi)^{-1/2} \tau_j^{-1} \left[\Phi\left[\left(\upsilon_j - \underline{\beta}_j\right) \Big/ \tau_j\right] - \Phi\left[\left(\lambda_j - \underline{\beta}_j\right) \Big/ \tau_j\right]\right]^{-1} \mathrm{I}_{(\lambda_j, \upsilon_j)}(\beta_j)$$

(where $b = \sum_{i=1}^{n} x_{ij} z_i \Big/ \sum_{i=1}^{n} x_{ij}^2$ and $\omega^2 = \sigma^2 \Big/ \sum_{i=1}^{n} x_{ij}^2$)

$$= \exp\left[-\sum_{i=1}^{n} \left(z_i - b x_{ij}\right)^2 \Big/ 2\sigma^2\right] \exp\left[-\left(\beta_j - \overline{\beta}_j\right)^2 \Big/ 2\sigma_*^2\right]$$
$$\cdot \exp\left[-\left(b^2 / 2\omega^2 + \underline{\beta}_j^2 / 2\tau_j^2 - \overline{\beta}_j^2 / 2\sigma_*^2\right)\right] \tag{2.6}$$
$$\cdot (2\pi)^{-1/2} \tau_j^{-1} \left[\Phi\left[\left(\upsilon_j - \underline{\beta}_j\right) \Big/ \tau_j\right] - \Phi\left[\left(\lambda_j - \underline{\beta}_j\right) \Big/ \tau_j\right]\right]^{-1} \mathrm{I}_{(\lambda_j, \upsilon_j)}(\beta_j)$$

(where $\sigma_*^2 = \left(\omega^{-2} + \tau_j^{-2}\right)^{-1}$ and $\overline{\beta}_j = \sigma_*^2 \left(\omega^{-2} b + \tau_j^{-2} \underline{\beta}_j\right)$).

If the normal prior distribution for β_j is not truncated (*i.e.*, $\lambda_j = -\infty$, $\upsilon_j = +\infty$) then conditional on $\beta_j \neq 0$, $\beta_j \sim \mathrm{N}\left(\overline{\beta}_j, \sigma_*^2\right)$ -- the standard result for a normal prior mean when variance is known. If the normal prior distribution is truncated, then conditional on $\beta_j \neq 0$, $\beta_j \sim \mathrm{N}\left(\overline{\beta}_j, \sigma_*^2\right)$ truncated to the interval $\left(\lambda_j, \upsilon_j\right)$, or

$$\beta_j \sim \mathrm{TN}_{(\lambda_j, \upsilon_j)}\left(\overline{\beta}_j, \sigma_*^2\right). \tag{2.7}$$

To remove the conditioning on $\beta_j = 0$ or $\beta_j \neq 0$ it is necessary to integrate (2.6) over β_j and compare this expression to (2.5) The integration yields

$$\begin{aligned} &= \exp\left[-\sum_{i=1}^{n}\left(z_i - bx_{ij}\right)^2 / 2\sigma^2\right] \exp\left[-\left(b^2/2\omega^2 + \underline{\beta}_j^2/2\tau_j^2 - \overline{\beta}_j^2/2\sigma_*^2\right)\right]\left(\sigma_*/\tau_j\right) \\ &\quad \cdot \left[\Phi\left[\left(\upsilon_j - \overline{\beta}_j\right)/\sigma_*\right] - \Phi\left[\left(\lambda_j - \overline{\beta}_j\right)/\sigma_*\right]\right]\left[\Phi\left[\left(\upsilon_j - \underline{\beta}_j\right)/\tau_j\right] - \Phi\left[\left(\lambda_j - \underline{\beta}_j\right)/\tau_j\right]\right]^{-1}. \end{aligned}$$

Thus the conditional Bayes factor in favor of $\beta_j \neq 0$, versus $\beta_j = 0$, is

$$\begin{aligned} BF &= \exp\left[\sum_{i=1}^{n} z_i^2 - \sum_{i=1}^{n}\left(z_i - bx_{ij}\right)^2 / 2\sigma^2\right] \exp\left[-\left(b^2/2\omega^2 + \underline{\beta}_j^2/2\tau_j^2 - \overline{\beta}_j^2/2\sigma_*^2\right)\right]\left(\sigma_*/\tau_j\right) \\ &\quad \cdot \left[\Phi\left[\left(\upsilon_j - \overline{\beta}_j\right)/\sigma_*\right] - \Phi\left[\left(\lambda_j - \overline{\beta}_j\right)/\sigma_*\right]\right]\left[\Phi\left[\left(\upsilon_j - \underline{\beta}_j\right)/\tau_j\right] - \Phi\left[\left(\lambda_j - \underline{\beta}_j\right)/\tau_j\right]\right]^{-1} \mathrm{I}_{(\lambda_j, \upsilon_j)}\left(\beta_j\right) \\ &= \exp\left[\overline{\beta}_j^2/2\sigma_*^2 - \underline{\beta}_j^2/2\tau_j^2\right]\left(\sigma_*/\tau_j\right) \\ &\quad \cdot \left[\Phi\left[\left(\upsilon_j - \overline{\beta}_j\right)/\sigma_*\right] - \Phi\left[\left(\lambda_j - \overline{\beta}_j\right)/\sigma_*\right]\right]\left[\Phi\left[\left(\upsilon_j - \underline{\beta}_j\right)/\tau_j\right] - \Phi\left[\left(\lambda_j - \underline{\beta}_j\right)/\tau_j\right]\right]^{-1}. \end{aligned} \tag{2.8}$$

To draw β_j from its conditional distribution the conditional posterior probability that $\beta_j = 0$ is computed from the conditional Bayes factor (2.8):

$$\overline{p}_j = \frac{\underline{p}_j}{\underline{p}_j + \left(1 - \underline{p}_j\right) BF}. \tag{2.9}$$

Based on a comparison of a drawing from the uniform distribution on [0, 1] with this probability, the choice $\beta_j = 0$ or $\beta_j \neq 0$ is made. If $\beta_j \neq 0$ then β_j is drawn from (2.7).

Conditional on all β_j,

$$\left[\underline{\upsilon}\underline{\sigma}^2 + (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)\right] / \sigma^2 \sim \chi^2(\underline{\nu} + n).$$

The Gibbs sampling computational algorithm proceeds in the usual way. After an initial value for (β, σ) is drawn from the prior distribution, the parameters $\beta_1, \beta_2, \ldots, \beta_k, \sigma$

are drawn from their respective conditional posterior distributions. In most applications a key objective is determination of the posterior probability of each of the $2^{(k-k^*)}$ models. This could be done in the obvious way, by recording an indicator variable for the model corresponding to the non-zero β_j at the end of each iteration. More accurate approximations may be based on (2.9), however. In each *step* of each iteration, record the value $\bar{p}_j$ for the model corresponding to $\beta_j = 0$ and $\beta_l (l \neq j)$ either zero or non-zero as is the case in the conditional distribution for β_j. Similarly record the value $(1 - \bar{p}_j)$ for the model corresponding to $\beta_j \neq 0$ and the values of $\beta_l (l \neq j)$. In similar fashion, posterior expectations of functions of interest of the parameter vector (β, σ) may be approximated more efficiently by drawing a value for $\beta_j \neq 0$ whether β_j is set to zero in the Markov chain or not, and weighting by the probability $\bar{p}_j$ for the function with $\beta_j = 0$ and by $(1 - \bar{p}_j)$ for the function with $\beta_j \neq 0$.

2.3 Special cases

Several special cases of the Bayes factor (2.8) are of interest, because they indicate certain aspects of the relation between the prior and posterior distributions.

If $\tau_j \to \infty$ then $BF \to 0$ unless λ_j and υ_j are both finite. This is a manifestation of Lindley's paradox (Bartlett, 1957; Lindley, 1957): as the prior distribution under one hypothesis becomes increasingly diffuse the posterior odds ratio in favor of the alternative increases without bound.

If $\underline{\beta}_j = 0$ and $\tau_j \to 0$, then $BF \to 1$, because the conditional distribution of β_j given $\beta_j \neq 0$ becomes identical with the conditional distribution of β_j given $\beta_j = 0$.

Finally, if $\underline{\beta}_j \neq 0$ and $\tau_j \to 0$, BF approaches

$$\exp\left\{\left[\sum_{i=1}^{n} z_i^2 - \sum_{i=1}^{n} \left(z_i - \underline{\beta}_j x_{ij}\right)^2\right] \Big/ 2\sigma^2\right\}.$$

In the limit the Bayes factor (2.8) compares two simple hypotheses and it becomes the ratio of the conditional likelihood functions given each of the hypotheses.

2.4 Computational efficiency

Since the Gibbs sampling algorithm described here employs complete blocking, the degree of serial correlation in the Monte Carlo Markov chain generated by the Gibbs sampler will depend on the degree of multicollinearity in the correlation matrix of the regressors (Geweke, 1991). If all sample correlation coefficients were zero, then the draw from the Gibbs sampler would be serially uncorrelated. However such a situation would be exceptional. Indeed, the most interesting and difficult cases -- the ones for which

proceeding to a formal analysis of the kind described here is most compelling -- are precisely those in which there is a high degree of collinearity among regressors. Experience with the algorithm indicates that the higher the ratio of the largest to smallest eigenvalues of the sample correlation matrix of the regressors, the more iterations will be required to achieve the same degree of numerical accuracy. For small regression problems the algorithm is fast: using code for which little optimization has been undertaken, 10^6 iterations for a 5-regressor model requires about 40 seconds on a Sun 10/51 with untruncated normal priors and about 75 seconds with truncated normal priors. Execution time appears to be roughly proportional to the cube of the number of regressors, so computation time for larger models can be much longer.

3. Noncontingent variable selection: Examples

We now take up three specific examples of noncontingent variable selection in regression. The examples are the same ones considered in George and McCulloch (1993). The objectives in these examples are to demonstrate a convenient method for the formulation of subjective priors, illustrate the numerical accuracy of the procedure, and study the relation between prior and posterior distributions in this model.

3.1 The happiness data

These data were collected from 39 employed MBA students in a class at the University of Chicago Graduate School of Business. Five variables were recorded: y_i = Happiness, recorded on a 10-point scale with 1 representing a suicidal state, 5 a feeling of "just muddling along" and 10 a euphoric state; x_{i1} = Money, measured by family income in thousands of dollars; x_{i2} = Sex, measured by 0 or 1 with 1 being a satisfactory level of sexual activity; x_{i3} = Love, with 1 indicating loneliness and isolation, 2 a set of secure relationships, and 3 a deep feeling of belonging and caring in a family or community; x_{i4} = Work, recorded on a 5-point scale with 1 indicating that the individual is seeking other employment, 3 that the individual's job is "OK", and 5 indicating that the job is enjoyable.

The linear regression model is

$$y_i = \beta_1 + \sum_{j=1}^{4} \beta_{j+1} x_{ij} + \varepsilon_i,$$

a specific instance of (2.1). Each regressor is a candidate for inclusion in the model, and there are several interesting hypotheses to be entertained about various combinations. George and McCulloch mention the hypotheses that Love and Work are the determinants of Happiness; readers may add others.

For this example, consider specification of a prior distribution that reflects the belief that each regressor may be a substantively significant determinant of Happiness, or it may not enter the model at all. We interpret "substantially significant determinant" to mean that a major change in the regressor in question, Δx_{ij}, ought to bring about a major change in the dependent variable, Δy_i. We then set the mean of the normal prior distribution to $\underline{\beta}_j = 0$ and take the prior standard deviation $\tau_j = \tau_j^* = \Delta y_i / \Delta x_{ij}$. For the results here major change in Happiness was set at $\Delta y_i = 4$; in Money at $\Delta x_{i1} = 50$; in Sex at $\Delta x_{i2} = .5$; in Love at $\Delta x_{i3} = 1$; and in Work at $\Delta x_{i4} = 2$. The mapping from "substantially significant determinant" to the τ_j is itself subjective, and we present results for $\tau_j = .5\tau_j^*$ and $\tau_j = 2\tau_j^*$ as well as for $\tau_j = \tau_j^*$. For the intercept term we choose $\underline{\beta}_1 = 0$ and $\tau_1 = 9$, reflecting uncertainty about the inclusion of various combinations of x_{ij} in the model. For the standard deviation of ε_i, $\underline{\sigma} = 2.5$, based on a prior mean of 5 for the standard deviation of y_i and a prior mean of .75 for the multiple correlation coefficient; $\underline{\nu} = .01$.

Different prior beliefs will, of course, lead to other choices for the τ_j. For instance, if it is thought that a certain regressor may either be a substantially insignificant determinant of Happiness or it may not enter the model at all, then the corresponding τ_j would be smaller and its value may be set employing the same kind of reasoning about marginal effects. As in George and McCulloch (1993, p. 883) the idea is to support β_j that are different from 0, but not so large as to dilute support for realistic values with support for unrealistically large values. The scaling procedures of Mitchell and Beauchamp (1988) accomplish much the same objective. The procedure followed here is exactly that suggested by Berger (1988).

For illustrative purposes we take as a base prior probability that each variable is excluded from the model, $\underline{p}_j = .5 \; (j = 2, \ldots, 5)$; the intercept β_1 is always included ($\underline{p}_1 = 0$). To study the relation between the prior and posterior distributions, we also consider $\underline{p}_j = .2$ $(j = 2, \ldots, 5)$ and $\underline{p}_j = .8 \; (j = 2, \ldots 5)$, always maintaining $\underline{p}_1 = 0$.

For each combination of the τ_j and $\underline{p}_j$ we consider half-normal as well as normal priors for the $\beta_j \; (j = 2, \ldots, 5)$. In the half-normal prior, $\lambda_j = 0$ and $\upsilon_j = \infty$, imposing a prior belief that these coefficients will be positive if the corresponding variable enters the model at all.

Summaries of the happiness data and prior distribution are provided in Table 1. Collinearity among regressors is modest. Posterior probabilities of alternative models corresponding to the normal priors are presented in Table 2, and the half-normal priors in Table 3. These results are obtained using the methods described in Section 2, with $m = 10^5$ iterations of the Gibbs sampling algorithm. The numerical accuracy of these results was assessed in two ways. The first uses the numerical standard error and relative numerical

efficiency discussed in Geweke (1991) for the Gibbs sampling algorithm. Relative numerical efficiency, in turn, may be expressed in terms of an i.i.d.-equivalent number of iterations, m^*: that is, the numerical accuracy achieved for m iterations of the Gibbs sampling algorithm is the same as that for m^* hypothetical i.i.d. drawings from the posterior. Thus, for a posterior probability $\overline{p}_j$ presented in Table 2 the numerical standard error is $\left[\overline{p}_j\left(1-\overline{p}_j\right)/m^*\right]^{1/2}$. In Tables 2 and 3, m^* ranges from 1,250 to 25,000 for the $.5\tau_j^*$ priors; from 960 to 24,000 for the τ_j^* priors; and from 570 to 24,000 for the $2\tau_j^*$ priors. The lower bound on numerical accuracy is higher for small values of the τ_j, because the smaller values reduce the collinearity in the posterior precision matrix for β, thereby reducing serial correlation in the Gibbs sampler and increasing computational efficiency.

The second method for assessing numerical accuracy is based on the observation that for the same values of the prior standard deviations τ_j, changing common values of the $\underline{p}_j$ $(j = 2,\mathsf{K}\ 5)$ from one common value to another will not affect the relative posterior probability of those models with the same number of regressors. For example, the posterior probability of all 2-variable models conditional on the model containing two variables and $\tau_j = \tau_j^*$ should be the same whether $\underline{p}_j = .2$, $\underline{p}_j = .5$ or $\underline{p}_j = .8$. These conditional probabilities -- simply the ratio of entries for specific models to the entry in the "2 regressors" row of the same columns of Tables 2 and 3 -- are indeed equal to within the tolerance indicated by the i.i.d.-equivalent number of iterations m^*.

Regardless of the particular prior distribution the models with regressors Love alone (x_{i3}), Love and Work (x_{i3} and x_{i4}) or Love, Work and Money (x_{i3}, x_{i4}, and x_{i1}) have total posterior probability at least two-thirds and often much more. Three systematic effects of the prior distributions on the posterior probabilities of the alternative models are evident. First, increases in $\underline{p}_j$, the prior probability that $\beta_j = 0$, favor smaller models. Second, increases in τ_j also favor smaller models. This is due to the fact that the values of the τ_j are all fairly large compared to the mass of the likelihood function (compare the τ_j^* and least squares coefficients in Table 1). As the τ_j increase over this range, the Bayes factors corresponding to models with large numbers of regressors decrease relative to those for models with smaller numbers of regressors, and the magnitude of the prior density decreases in the region of the mass of the likelihood function. In the limit, as all $\tau_j \to \infty$, all posterior probability becomes concentrated on the model with no regressors, consistent with Lindley's paradox. For these data and the range of τ_j's employed the effect is to move the model probability from the Love-Work-Money model to the Love-Work model to the Love model. Finally, for given values of the τ_j and $\underline{p}_j$ smaller models receive higher posterior probability under normal priors than they do under half-normal priors. This is

due to the fact that (except for Sex, which turns out to be unimportant) the bulk of the likelihood function is in the positive orthant. Since the half-normal priors concentrate their probability there, they favor larger models more than do the normal priors.

Our results are only indirectly comparable with those of George and McCulloch (1993). Even the likelihood function is not the same, since (as discussed in the introduction) they do not account for uncertainty about the intercept. However, results are broadly consistent. Both their results and ours favor the Love model among all one-variable models, the Love-Work model among all two-variable models, and the Love-Work-Money model among all three-variable models. The approach taken here has some tendency to favor smaller models than does the George-McCulloch approach, but differences are not great.

3.2 The Hald data

These data are presented in Draper and Smith (1981) and are often used to illustrate techniques for selective regressors. The five variables are y_i = Heat produced in the hardening of cement (in calories per gram), x_{i1} = percentage of input composed of tricalcium aluminate, x_{i2} = percentage of input composed of tricalcium silicate, x_{i3} =percentage of input composed of tetracalcium alumino ferrite, and x_{i4} = percentage of input composed of dicalcium silicate. There are only 13 observations in the data set.

The prior distributions were constructed in the same way as in the previous example, taking $\Delta y_i = 20°$ as a major change in the dependent variable. Major changes in regressor variables were set to one-half their range in the data set: $\Delta x_{i1} = 10$, $\Delta x_{i2} = 22.5$, $\Delta x_{i3} = 8.5$, and $\Delta x_{i4} = 27$. Alternative values of the τ_j were set accordingly and once again either $\underline{p}_j = .2$ or $\underline{p}_j = .5$ or $\underline{p}_j = .8$ for all $j = 2, \mathsf{K}, 5$. A summary of the τ_j^* and of the data is provided in Table 4. Collinearity in the regressors is quite high, with the ratio of largest to smallest eigenvalues of the correlation matrix exceeding 10^3.

Posterior probabilities for alternative models are presented in Tables 5 and 6 in the same format as were the results for the Happiness data in Tables 2 and 3. In view of the ill-conditioned data, computations employed 10^6 iterations of the Gibbs sampling algorithm, 10 times more than for the Happiness data. The number of i.i.d.-equivalent iterations m^*, for purposes of judging numerical accuracy, ranged from $m^* = 2,000$ to $m^* = 15,000$ for priors with $\tau_j = .5\tau_j^*$, from $m^* = 580$ to $m^* = 13,000$ for priors with $\tau_j = \tau_j^*$, and from $m^* = 30$ to $m^* = 6,000$ when $\tau_j = 2\tau_j^*$. The explanation for the effect of τ_j^* on m^* is the same, but the magnitude is larger for the Hald data because of the greater collineariety of the covariates.

Overall the favored models incorporate x_{i1} and x_{i2}; x_{i1}, x_{i2} and x_{i3}; x_{i1}, x_{i2}, and x_{i4}; or all four regressors. Between them these four models always account for at least 80% of the posterior probability, and in some cases close to 100%. The qualitative effects of changing the $\underline{p}_j$ or the τ_j, or of moving from normal to half-normal prior distributions, are the same as in the Happiness data for the same reasons. However the sensitivity of the posterior model probabilities to changes in the prior is much greater, as one would expect both from the conditioning of the regressor moment matrix and the small sample size.

George and McCulloch also find posterior model probabilities sensitive to their prior distribution., but in most other respects their results differ from ours. In general, their methods produce models with small numbers of regressors, producing probability .44 of no regressors in one case, and never producing a probability greater than .02 for the model with all four regressors. Their most probable two-variable model is the same as ours. They are unable to discriminate among the three-regressor models (x_{i1}, x_{i2}, x_{i3}), (x_{i1}, x_{i2}, x_{i4}) and (x_{i1}, x_{i2}, x_{i4}), whereas the results in Tables 5 and 6 place relatively less posterior probability on (x_{i1}, x_{i2}, x_{i4}).

3.3 The Bank data

The third data set was collected from 233 branches of a major New York bank to compare sales of new accounts with bank characteristics. The dependent variable is the number of new accounts sold in a particular time period. The explanatory variables are listed in Table 7, together with their "major change" definitions used to construct the prior parameters τ_j^* when combined with the specification that 400 constitutes an appreciable increase in the number of new accounts. The sample correlation matrix of the covariates and its eigenvalues are displayed in Table 8. Although this model has about four times the number of regressors as the Hald data model, the range of eigenvalues is similar.

Here we confine detailed discussion to results for a single prior, with $\tau_j = \tau_j^*$ and $\underline{p} = .5$. The Gibbs sampling algorithm employed 2×10^6 iterations and produced numerical approximations whose accuracy is about the same as would have been achieved by $m^* = 2,000$ hypothetical i.i.d. drawings directly from the posterior distribution. Repetition of the procedure with different seeds yields the same results as those in Tables 9 and 10 up to differences consistent with this degree of numerical accuracy. Table 9 presents the marginal posterior probability of inclusion in the model for each variable, and compares it with the "p-value" from the ordinary least squares regression. In general posterior inclusion probabilities and p-values are inversely related, but variables 1 and 3 constitute exceptions.

Table 10 shows those 20 models whose posterior probabilities exceed .01; the posterior probability of these 20 models jointly is .4944. Regressors 6, 8, 9 and 14 appear in all 20 models; regressor 12 never appears. Thus 10 of the 15 regressors appear in some but not all models. Using the same data set but priors of a different type, George and McCulloch (1993) report that only 10 models have posterior probability above .01. The regressors common to all the models in Table 10 are also common to all the George-McCulloch models, but in addition their models always include regressors 1, 7, 10, 13 and 15. Overall, the George-McCulloch models are generally larger than those reported here and there is less variety in the mix of regressors in their models. None of the models in Table 10 receive posterior probability above .01 with the George-McCulloch priors. Moreover, changing the prior distribution in the same was as was done in the previous two examples substantially affects the posterior distribution: typically only 3 to 7 of the models appearing in Table 10 will appear in the corresponding table for the new posterior distribution.

4. Contingent variable selection

In the prior distributions used to the point the probability that a coefficient is zero is unaffected by the values of the other coefficients. For the examples taken up in the previous section this is a natural and simple assumption but in other situations it need not be. The model selection problem is often framed as choosing one from an ordered or semi-ordered sequence of models. Some leading examples arise in time series when lag length must be chosen for a univariate autoregression (an ordering) or lag lengths is one equation from a vector autoregression (a semi-ordering). More generally when variables enter a model through a family of parametric expansions the problem of choosing the order of the model conditional on some maximum order has essentially the same statistical structure.

The basic method described in Section 2 can easily be modified to provide a workable Bayesian solution to the problem. Let $G_j = \left\{g_1^j, g_2^j, \mathsf{K}\ , g_{k_j}^j\right\}$ be a set of integers indicating those coefficients that must be nonzero if β_j is to be nonzero: *i.e.*, $\beta_j \neq 0 \Rightarrow \mathrm{P}\left[\beta_{g_i^j} \neq 0 \left(i = 1, \mathsf{K}\ , k_j\right)\right] = 1$. Let $\underline{p}_j^*$ denote the prior probability that $\beta_j = 0$ *conditional* on $\beta_{g_i^j} \neq 0\ \left(i = 1, \mathsf{K}\ , k_j\right)$. In the case of a complete ordering of the coefficients $\beta_1, \mathsf{K}\ , \beta_k$ the probability that exactly s coefficients are nonzero is simply $\prod_{j=1}^{s}\left(1 - \underline{p}_j^*\right)$. In the case of a semi-ordering arising from complete orderings in d dimensions (as in one equation from a vector autoregression) the prior probability of exactly s_q nonzero coefficients in dimension

$q\,(q=1,\mathsf{K}\,,d)$ is $\prod_{q=1}^{d}\prod_{j=1}^{s_q}\left(1-\underline{p}_{q_j}^{*}\right)$, where β_{q_j} is the s_q'th coefficient in dimension q. Parameterizing the prior distribution through the $\underline{p}_j^{*}$ is parsimonious, and -- as will be seen in the next section -- is often a convenient way to express prior beliefs.

The remainder of the prior distribution for the contingent variable selection problem is the same as for the noncontingent selection problem described in Section 2. Corresponding to each coefficient is a proper normal prior distribution, possibly truncated to the finite or semi-infinite interval $\left(\lambda_j,\upsilon_j\right)$.

Conditional posterior distributions may be developed closely following Section 2.2, leading to the Bayes factor (2.8) in favor of $\beta_j\neq 0$ versus $\beta_j=0$, conditional on $\beta_{\mathsf{l}}(\mathsf{l}\neq j)$; the posterior probability that $\beta_j\neq 0$ conditional on $\beta_{\mathsf{l}}(\mathsf{l}\neq j)$, which is (2.9) with $\underline{p}_j$ replaced by $\underline{p}_j^{*}$; and finally, if $\beta_j\neq 0$ the posterior distribution of β_j given by (2.7). Thus once the table of integers $G_j\,(j=1,\mathsf{K}\,,k)$ is developed, only minor changes in the computational algorithm are required to draw from the posterior distribution for contingent variable selection using a Gibbs sampling algorithm.

Convergence of the Gibbs sampling algorithm rests on the irreducibility of the continuous state space Markov chain inherent in the algorithm, which in turn may be discerned from the $G_j\,(j=1,\mathsf{K}\,,k)$. In general verifying irreducibility could be tedious. In the case of a complete ordering it is trivial: $\underline{p}_j^{*}<1\forall j$ is sufficient, for then there is positive probability of reaching any open neighborhood in the support of the posterior distribution from any point, in k or fewer steps. The same is true for a semi-ordering arising from complete orderings in d dimensions, except that $\sup_{q=1,\mathsf{K}\,,d}s_q$ or fewer steps are required.

5. Contingent variable selection: an example

This section presents a single example of contingent variable selection. (Further examples will be added in a revised version of this paper.) The example consists of a univariate autoregression

$$y_t=\beta_1+\sum_{j=1}^{8}\beta_{j+1}y_{t-j}+\varepsilon_t\quad(t=1,\mathsf{K}\,,T),$$

where y_t denotes the logarithm of annual real gross national product in the United States, $t=1$ corresponds to 1918, and $t=T$ corresponds to 1970. The data are taken from the influential study of Nelson and Plosser (1982) on trend and difference stationarity in macroeconomic time series.

The sample correlation matrix of the covariates is provided in Table 11. As one would expect for these data, the matrix is ill conditioned. The ratio of largest to smallest eigenvalue is about the same as for the Hald data, but here there is a dominant eigenvalue corresponding to the trend in the log real GNP data.

The prior distribution for the β_j conditional on $\beta_j \neq 0$, is very similar to that suggested by Doan, Litterman and Sims (1984). The β_j are independently distributed, with

$$\beta_2 \sim \mathrm{N}(1,\ \pi_0\pi_1), \quad \beta_j \sim \mathrm{N}\left(0,\ \pi_0\pi_1^{(j-1)}\right) \quad (j = 3,\mathrm{K}\ ,9).$$

For β_1, which always enters the model, the diffuse prior $\beta_1 \sim \mathrm{N}\left(0, 10^{20}\right)$ was employed; alternative priors for β_1 have little effect on variable selection.

Two prior distributions were selected for illustration. In the first, π_0 and π_1 were chosen so that $\beta_2 \sim \mathrm{N}\left(1, .5^2\right)$ and $\beta_9 \sim \mathrm{N}\left(0, .1^2\right)$, which means $\pi_1 = .1822$. The prior parameters $\underline{p}_j^* = .1822$ also. Thus, the prior distribution reflects the belief that coefficients of longer lags, if they are nonzero, are likely to be smaller in absolute value than coefficients on shorter lags. This is the belief that led to the functional form chosen by Doan, Litterman and Sims (1984). The prior distribution also reflects the belief that, conditional on a maximum lag length of 8 years, shorter lag lengths are more likely than larger ones, with a lag length of 8 only 1/5 as probably as a lag length of zero. In the second prior distribution π_0 and π_1 were chosen so that $\beta_2 \sim \mathrm{N}\left(1, 1^2\right)$ and $\beta_9 \sim \mathrm{N}\left(0, .1^2\right)$; $\pi_1 = .2500$ and $\underline{p}_j^* = .2500$. This second prior distribution favors shorter lag lengths more than the first, both because of the higher prior conditional probability that any coefficient is zero, and also because the prior distribution for nonzero coefficients is more diffuse.

Serial correlation in the Gibbs sampler was quite substantial for this example, so 10^6 iterations were employed. (Execution time was about 17 minutes on a Sun 10/51.) Measuring accuracy in the same way as for the examples in Section 3, these 10^6 iterations produced results whose numerical accuracy is the same as if about 300 i.i.d. drawings from the posterior distribution had been made. For each lag, Table 12 indicates the familiar Akaike Information Criterion (AIC) and Schwarz Bayesian Information Criterion (SBIC), along with the probability that the lag enters the model ("P(Lag)") and the probability that the lag is last, in which case it is the order of the model ("P(Model)"). The modal model order is 2 lags for both priors, which coincides with the model that minimizes both the AIC and SBIC model choice criterion. As anticipated, the first prior distribution leads to longer lags than the second prior distribution, but the differences are not great: e.g., one-sided 75%, 90% and 95% confidence intervals for model size are the same for each posterior distribution.

6. Summary and Extensions

This paper has proposed a family of nonconjugate priors for subjective Bayesian treatment of variable selection and model comparison in linear regression. Since priors for different coefficients are independent the investigator can consider one coefficient at a time. This investigator is, however, forced to think explicitly about plausible magnitudes for each coefficient conditional on the corresponding regressor appearing in the model. Sign restrictions and other limitations on support are easily accommodated.

In experiments with two models having five regressors each, it was found that priors interact with data in an understandable way: *e.g.*, increased probability of variable exclusion leads to smaller models, as do increasingly diffuse priors for coefficients of included variables. The computational efficiency of the Gibbs sampling computational algorithm is largely a function of collinearity in the posterior distribution of the coefficients: *e.g.*, the more ill-conditioned the covariate sample correlation matrix the less efficient the algorithm; the greater the prior precision of coefficients of included variable the more efficient the algorithm.

The problem of Bayesian choice of regressors as typically set forth (e.g., by Lempers (1971), Stewart (1987), Mitchell and Beauchamp (1988), and George and McCulloch (1993)) was extended in this paper to arbitrarily complex sets of contingencies for variable entry into the model. This extension makes it possible to take up the task of lag length selection in time series, and the problem of model selection from parametric expansion families like Taylor and Laurent series. In all the examples considered the posterior distribution changes in important ways in response to reasonable changes in the prior distribution. This fact underscores the importance of choosing prior distributions carefully, and should make subjective Bayesians even more wary of "automatic" procedures that seek to avoid explicit specification of priors. Only in small models with many observations -- many more than employed in the examples taken up here -- will the posterior be robust to reasonable changes in the prior. But given many observations investigators generally enlarge the number of models considered, thus perpetuating sensitivity to the prior distribution.

Several extensions to these developments are natural and would involve no problems beyond normal technical difficulties in implementation. Following West (1984), Geweke (1993), Diebold and Robert (1994), and others, the assumption that disturbances are normal may be weakened through appropriate use of mixture models. Nor is the procedure limited to truncated normal prior distributions for coefficients of included variables: since the Gibbs sampling algorithm is fully blocked essentially arbitrary prior distributions may be specified

for each coefficient with no serious technical impediment. Extension of the methods proposed here to multivariate regression models in general and vector autoregressions in particular is also straightforward. Finally, through modest modification of the fully blocked Gibbs sampling algorithm introduced in Geweke (1994) one could adapt the algorithm of this paper to variable selection and sign restrictions in essentially any model for which the posterior distribution can be expressed in closed form; whether the algorithm would still be fast enough to be practical is an open question.

References

Bartlett, M.S., 1957, "A Comment on D.V. Lindley's Statistical Paradox," *Biometrika* **44**: 533-534.

Berger, J.O. 1988, "Comment" (on Mitchell and Beauchamp, 1988), *Journal of the American Statistical Association* **83**: 1033-1034.

Diebold and Robert -- current paper in JRSS

Doan, T., R. Litterman and C. A. Sims, 1984, "Forecasting and Conditional Projection Using Realistic Prior Distributions," *Econometric Reviews* **5**(1): 1-52.

Draper, N., and H. Smith, 1981, *Applied Regression Analysis* (Second edition). New York: John Wiley.

George, E.I. and R.E. McCulloch, 1993, "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association* **88**: 881-889.

Geweke, J., 1991, "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments," in J.O. Berger, J.M. Bernardo, A.P. Dawid, and A.F.M. Smith (eds.), *Proceedings of the Fourth Valencia International Meeting on Bayesian Statistics*, 169-194. Oxford: Oxford University Press, 1992.

Geweke -- JAE paper

Geweke, J., 1994, "A Computational Engine for Bayesian Inference in Econometric Models," Federal Reserve Bank of Minneapolis working paper.

Lempers, R.B., 1971, *Posterior Probabilities of Alternative Linear Models*. Rotterdam: Rotterdam University Press.

Lindley, D.V., 1957, "A Statistical Paradox," *Biometrika* **44**: 187-192.

Miller, A.J., 1990, *Subset Selection in Regression.* New York: Chapman and Hall.

Mitchell, T.J. and J.J. Beauchamp, 1986, "Algorithms for Bayesian Variable Selection in Regression," in T. J. Boardman (ed.), *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface*. Washington: American Statistical Association.

Mitchell, T.J. and J.J. Beauchamp, 1988, "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association* **83**: 1023-1032.

Nelson, C.R., and C. I. Plosser, 1982. "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications," *Journal of Monetary Economics* **10**: 139-162.

Poirier, D.J., 1985, "Bayesian Hypothesis Testing in Linear Models with Continuously Induced Conjugate Priors Across Hypotheses," in J.M. Bernardo et al. (eds.), *Bayesian Statistics 2*, pp. 711-722. Amsterdam: North-Holland.

Stewart, L., 1987, "Hierarchical Bayesian Analysis using Monte Carlo Integration: Computing Posterior Distributions when there are Many Possible Models," *The Statistician* **36**: 211-219.

Tierney, L., 1991. "Markov Chains for Exploring Posterior Distributions." University of Minnesota School of Statistics Technical Report No. 560.

West, M., 1984. "Outlier Models and Prior Distributions in Bayesian Linear Regression," *Journal of the Royal Statistical Society Series B* **46**: 431-439.

Table 1

Happiness data

Variable	Definition	Least squares coefficient	Least squares standard error	Prior standard deviation τ_j^*
1	Money	.00958	.00521	.08
2	Sex	-.149	.419	8.00
3	Love	1.92	.295	4.00
4	Work	.476	.199	2.00

Covariate sample correlation matrix,

1.000	.307	.126	.068
	1.000	.047	-.316
		1.000	.386
			1.000

Eigenvalues of covariate sample correlation matrix

.4405	.7356	1.3468	1.477

Table 2

Model posterior probabilities, Happiness data with full normal priors

$\underline{p}_j = P(\beta_j = 0)$	---------- 0.2 ----------			---------- 0.5 ----------			---------- 0.8 ----------		
τ_j	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$
0 regressors									
1									
2									
3	.0074	.0257	.0789	.0767	.1704	.3513	.3387	.4948	.7077
4				.0001		.0001	.0003		
1 regressor	.0074	.0257	.0789	.0786	.1704	.3515	.3390	.4948	.7077
1,2									
1,3	.0186	.0310	.0459	.0469	.0499	.0521	.0505	.0387	.0255
1,4									
2,3							.0001		
3,4	.1822	.3289	.4668	.4699	.5454	.4831	.4943	.4170	.2488
2 regressors	.2049	.3662	.5229	.5262	.6071	.5466	.5557	.4636	.2800
1,2,3	.0174	.0150	.0113	.0112	.0060	.0037	.0029	.0010	.0004
1,3,4	.4780	.4266	.3015	.3019	.1777	.0831	.0868	.0343	.0102
1,2,4									
2,3,4	.0770	.0695	.0505	.0501	.0290	.0131	.0131	.0057	.0016
3 regressors	.5725	.5111	.3633	.3632	.2127	.0998	.1028	.0410	.0122
1,2,3,4	.2152	.0970	.0348	.0337	.0098	.0020	.0025	.0006	

All models have strictly positive posterior probability. Empty cells indicate posterior probability less than 10^{-4}.

Table 3

Model posterior probabilities, Happiness data with half-normal priors

$\underset{\sim}{p}_j = P(\beta_j = 0)$	---------- 0.2 ----------			---------- 0.5 ----------			---------- 0.8 ----------		
τ_j	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$
0 regressors									
1									
2									
3	.0025	.0098	.0317	.0271	.0893	.1867	.1763	.3394	.5368
4									
1 regressor	.0025	.0098	.0317	.0271	.0893	.1867	.1763	.3394	.5368
1,2									
1,3	.0013	.0247	.0352	.0315	.0552	.0574	.0554	.0455	.0377
1,4									
2,3	.0007	.0013	.0021	.0018	.0027	.0033	.0033	.0027	.0021
2,4									
3,4	.1230	.2561	.4318	.4130	.5556	.6212	.6309	.5693	.4296
2 regressors	.1367	.2561	.4138	.4130	.5556	.6212	.6309	.5693	.4296
1,2,3	.0024	.0024	.0015	.0014	.0013	.0007	.0007	.0003	.0001
1,3,4	.6079	.5815	.4691	.4709	.3109	.1706	.1708	.0812	.0302
1,2,4									
2,3,4	.0617	.0591	.0048	.0494	.0309	.0173	.0178	.0089	.0032
3 regressors	.6721	.6431	.5175	.5217	.3431	.1886	.1892	.0905	.0334
1,2,3,4	.1887	.0911	.0370	.0381	.0120	.0034	.0036	.0008	.0001

All models have strictly positive posterior probability. Empty cells indicate posterior probability less than 10^{-4}.

Table 4

Hald data

Variable	Definition	Least squares coefficient	Least squares standard error	Prior standard deviation τ_j^*
1	% Tricalcium aluminate	1.55	.745	2.0
2	% Tricalcium silicate	.510	.724	.89
3	% Tetracalcium alumino ferrite	.102	.755	2.1
4	% dicalcium silicate	-.144	.709	.74

Moment correlation matrix

1.000	.229	-.824	-.245
	1.000	-.139	-.973
		1.000	.030
			1.000

Eigenvalues of covariate sample correlation matrix

.001624	.1866	1.576	2.236

Table 5

Model posterior probabilities, Hald data with full normal priors

$\underline{p}_j = P(\beta_j = 0)$	---------- 0.2 ----------			---------- 0.5 ----------			---------- 0.8 ----------		
τ_j	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$
0 regressors									
1									
2									
3									
4									
1 regressor									.0001
1,2	.0936	.2018	.2882	.4468	.5881	.5834	.8007	.8589	.7469
1.3									
1.4		.0043	.0497		.0146	.1317	.0001	.0084	.1660
2,4									
3,4			.0004			.0009			.0008
2 regressors	.0936	.2061	.3383	.4468	.6028	.7161	.8008	.8673	.9137
1,2,3	.1969	.1873	.1301	.2306	.1359	.0062	.1079	.0505	.0212
1,3,4	.0001	.0065	.0052		.0057	.0381	.	.0008	.0119
1,2,4	.1504	.2748	.3097	.1746	.1967	.1605	.0749	.0755	.0514
2,3,4			.0004			.0001			
3 regressors	.3474	.4687	.4954	.4052	.3383	.2648	.1828	.1268	.0845
1,2,3,4	.5590	.3252	.1664	.1480	.0589	.0191	.0165	.0059	.0017

All models have strictly positive posterior probability. Empty cells indicate posterior probability less than 10^{-4}.

Table 6

Model posterior probabilities, Hald data with half-normal priors

$\underset{\sim}{p}_j = P(\beta_j = 0)$	---------- 0.2 ----------			---------- 0.5 ----------			---------- 0.8 ----------		
τ_j	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$
0 regressors									
1									
2									
3									
4									
1 regressor									.0008
1,2	.0406	.1325	.3002	.2842	.5411	.7647	.7110	.8618	.9363
1,3									
1,4									
2,3									
2,4									
3,4									
2 regressors	.0406	.1325	.3002	.2842	.5412	.7647	.7110	.8618	.9363
1,2,3	.1512	.2242	.2441	.2719	.2299	.1591	.1713	.9024	.0479
1.3.4									
1.2.4	.0648	.0750	.0798	.1164	.0782	.0464	.0682	.0323	.0139
2,3,4									
4 regressors	.2160	.2992	.3240	.3883	.3080	.2056	.2396	.1248	.0618
1,2,3,4	.7434	.5683	.3758	.3275	.1508	.0297	.0494	.0133	.0018

All models have strictly positive posterior probability. Empty cells indicate posterior probability less than 10^{-4}.

Table 7

Variable definitions and prior, Bank data

Varaible	Definition	Least squares coefficient	Least squares standard error	Major change, Δx_{ij}	Prior standard deviation τ_j
1	Number of households serviced	-.0153	.00498	8,000	.05
2	Number of people selling new account	11.5	4.06	4	100
3	Manhattan dummy	.48.2	23.7	5	800
4	Burroughs dummy	15.4	22.9	.5	800
5	Suburbs dummy	-8.40	21.6	.5	800
6	Demand deposits balance	-.0199	.00431	16,000	.025
7	Number of demand deposits	.0254	.00953	4,000	.1
8	NOW accounts balance	-.0131	.00437	10,000	.04
9	Number of NOW accounts	.544	.0524	600	.667
10	Money market accounts balance	-.00299	.00157	48,000	.008
11	Number of money market accounts	.0923	.0437	1200	.33
12	Passbook savings balance	.00976	.00248	8,000	.05
13	Other time deposit balance	.00568	.00220	24,000	.017
14	Consumer loans	.363	.0442	600	.667
15	Home mortgages	.00802	.00257	12,000	.033

Table 8

Covariate sample correlation matrix, Bank data

1.000	.739	.499	-.077	-.391	.812	.932	.587	.650	.672	.720	.501	.659	.853	.552
	1.000	.565	-.140	-.389	.702	.707	.647	.715	.691	.677	.214	.575	.769	.462
		1.000	-.411	-.541	.510	.504	.513	.414	.354	.287	.028	.222	.505	.314
			1.000	-.487	-.180	-.154	-.183	-.255	-.216	-.221	.217	-.137	-.190	-.132
				1.000	-.311	-.324	-.315	-.185	-.159	-.098	-.213	-.116	-.291	-.163
					1.000	.864	.844	.783	.834	.780	.351	.763	.828	.485
						1.000	.604	.670	.664	.720	.387	.616	.870	.587
							1.000	.831	.839	.710	.254	.698	.570	.250
								1.000	.899	.884	.286	.797	.678	.335
									1.000	.945	.329	.896	.681	.302
										1.000	.415	.907	.704	.370
											1.000	.564	.257	.150
												1.000	.648	.290
													1.000	.592
														1.000

Eigenvalues of covariate sample correlation matrix

.0163	.0288	.0321	.0455	.0623	.1154	.1663	.2290	.3506	.4073	.6676	1.067	1.527	1.692	8.593

Table 9

Marginal posterior probabilities of inclusion of variables, Bank data

Varaible	$P(\beta_j = 0)$	"p value"
1	.417	.002
2	.667	.005
3	.896	.043
4	.163	.502
5	.185	.698
6	.998	.000
7	.523	.008
8	.940	.003
9	1.000	.000
10	.399	.146
11	.555	.036
12	.085	.694
13	.761	.011
14	1.000	.000
15	.912	.002

Table 10

Some individual model posterior probabilities, Bank data

Inclusion and exclusion of regressors															Posterior probability
1	2	3	x	x	6	7	8	9	x	x	x	13	14	15	.1188
x	2	3	x	x	6	x	8	9	10	11	x	13	14	15	.0516
x	2	3	x	x	6	7	8	9	x	x	x	13	14	15	.0434
x	2	3	x	x	6	x	8	9	x	x	x	13	14	15	.0325
1	2	3	4	x	6	7	8	9	x	x	x	13	14	15	.0234
x	x	3	x	x	6	7	8	9	x	x	x	13	14	15	.0234
x	x	3	x	x	6	x	8	9	10	11	x	13	14	15	.0192
1	2	3	x	5	6	7	8	9	x	x	x	13	14	15	.0191
1	2	3	x	x	6	7	8	9	10	x	x	13	14	15	.0180
x	x	3	x	x	6	x	8	9	x	11	x	x	14	15	.0168
x	x	3	x	x	6	x	8	9	10	11	x	x	14	15	.0149
1	2	3	x	x	6	7	8	9	x	11	x	13	14	15	.0147
x	2	3	x	x	6	x	8	9	10	11	x	x	14	15	.0144
1	2	3	x	x	6	7	8	9	10	11	x	13	14	15	.0128
x	2	3	x	x	6	x	8	9	x	11	x	13	14	15	.0127
1	2	x	x	5	6	7	8	9	x	x	x	13	14	15	.0126
x	x	3	4	x	6	x	8	9	x	11	x	x	14	15	.0122
x	2	3	x	x	6	x	8	9	x	11	x	x	14	15	.0115
1	x	3	x	x	6	7	8	9	x	x	x	13	14	15	.0114
x	2	3	x	x	6	x	8	9	10	11	x	13	14	x	.0110

Table 11

Covariate sample correlation matrix, Real GNP data

1.000	.992	.977	.962	.949	.940	.935	.930
	1.000	.991	.977	.961	.947	.939	.933
		1.000	.991	.976	.959	.946	.937
			1.000	.991	.975	.958	.944
				1.000	.990	.974	.957
					1.000	.990	.973
						1.000	.990
							1.000

Eigenvalues of covariate sample correlation matrix

.0025	.0027	.0041	.0063	.0189	.0627	.1563	7.7464

Table 12

Posterior lag and model probabilities, Real GNP data

			---------Prior 1----------		---------Prior 2----------	
Lag	AIC	SBIC	P(Lag)	P(Model)	P(Lag)	P(Model)
1	-5.324	-5.249	1.0000	.0516	1.0000	.1286
2	-5.420	-5.309	.9484	.4310	.8714	.4878
3	-5.387	-5.238	.5174	.2869	.3836	.2246
4	-5.392	-5.206	.2305	.1393	.1590	.0592
5	-5.372	-5.149	.0912	.0519	.0638	.0420
6	-5.340	-5.070	.0393	.0218	.0218	.0153
7	-5.311	-5.013	.0175	.0098	.0065	.0032
8	-5.274	-4.939	.0077	.0077	.0033	.0033