Variable Selection Via Gibbs Sampling
Author(s): Edward I. George and Robert E. McCulloch
Source: *Journal of the American Statistical Association*, Vol. 88, No. 423 (Sep., 1993), pp. 881-889
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: http://www.jstor.org/stable/2290777
Accessed: 03-06-2018 21:38 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://about.jstor.org/terms

American Statistical Association, Taylor & Francis, Ltd. are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association*

This content downloaded from 98.249.87.104 on Sun, 03 Jun 2018 21:38:20 UTC
All use subject to http://about.jstor.org/terms

Variable Selection Via Gibbs Sampling

EDWARD I. GEORGE and ROBERT E. McCULLOCH*

A crucial problem in building a multiple regression model is the selection of predictors to include. The main thrust of this article is to propose and develop a procedure that uses probabilistic considerations for selecting promising subsets. This procedure entails embedding the regression setup in a hierarchical normal mixture model where latent variables are used to identify subset choices. In this framework the promising subsets of predictors can be identified as those with higher posterior probability. The computational burden is then alleviated by using the Gibbs sampler to indirectly sample from this multinomial posterior distribution on the set of possible subset choices. Those subsets with higher probability—the promising ones—can then be identified by their more frequent appearance in the Gibbs sample.

KEY WORDS: Data augmentation; Hierarchical Bayes; Latent variables; Mixture; Multiple regression.

A crucial problem in building a multiple regression model is the selection of predictors to include. More precisely, given a dependent variable Y and a set of potential predictors $X_1, \ldots, X_p$, the problem is to find and fit the "best" model of the form $Y = X_1^*\beta_1^* + \cdots + X_q^*\beta_q^* + \varepsilon$, where $X_1^*, \ldots, X_q^*$ is a "selected" subset of $X_1, \ldots, X_p$. A wide variety of selection procedures based on a comparison of all 2^p possible submodels have been proposed, including AIC, Cp, and BIC. Unfortunately, when p is large, the computational requirements for these procedures can be prohibitive. To mitigate the computational issue, practitioners typically use heuristic methods to restrict attention to a smaller number of potential subsets. This is the idea behind, for example, stepwise procedures, such as forward selection or backward elimination, which sequentially include or exclude variables based on R^2 considerations. Miller (1990) provided a comprehensive summary and bibliography of these procedures.

The main thrust of this article is to develop a procedure that we call SSVS (stochastic search variable selection) to select "promising" subsets of $X_1, \ldots, X_p$ for further consideration. SSVS is based on embedding the entire regression setup in a hierarchical Bayes normal mixture model, where latent variables are used to identify subset choices. In this framework the promising subsets of predictors can be identified as those with higher posterior probability. SSVS then proceeds by using Gibbs sampling to indirectly sample from this multinomial posterior distribution on the set of possible subset choices. Those subsets with higher probability—the promising ones—can then be identified by their more frequent appearance in the Gibbs sample. In this way SSVS avoids the overwhelming problem of calculating the posterior probabilities of all 2^p subsets.

SSVS is controlled by various tuning parameters that can be prespecified by the user. With different prespecifications, the user can address the particular goals of variable selection that are appropriate for the problem under consideration. Such goals may include, for example, the search for a parsimonious model that does not drastically increase the error of approximation or the elimination of ensembles of variables that are unimportant compared to their sampling uncertainty. A distinguishing feature of SSVS is that it allows the user to let the practical importance of a variable influence its selection, rather than just its statistical significance.

The background literature for SSVS is based on two lines of research. First, our use of a hierarchical Bayes model to identify the "promising" variables has its roots in the literature on Bayesian model discrimination. Some of the work related to our approach includes Lempers (1971), Atkinson (1978), Perrichi (1984), Smith and Spiegelhalter (1980), Spiegelhalter and Smith (1982), Zellner (1984), Poirier (1985), Stewart (1987), and especially Mitchell and Beauchamp (1988). The second line of background research concerns Gibbs sampling; see Casella and George (1992) for an elementary introduction. Papers particularly relevant to our use of Gibbs sampling include Diebolt and Robert (in press), Gelfand and Smith (1990), Gelfand, Hills, Racine-Poon, and Smith (1990), Tanner and Wong (1987), and Verdinelli and Wasserman (1991).

The plan of this article is as follows. In Section 1 we define and motivate the hierarchical framework that serves as the basis for SSVS. In Section 2 we show how this hierarchical model can be used to identify the most promising regression models, and Section 3 we show how SSVS via the Gibbs sampler can efficiently identify these promising subsets. In Section 4 we illustrate SSVS on simulated examples, and in Section 5 we apply SSVS to real data sets.

1. A HIERARCHICAL MODEL FOR VARIABLE SELECTION

For the regression situation involving the observation of a dependent variable Y and a set of potential predictors $X_1, \ldots, X_p$, we consider the canonical regression setup

$$\mathbf{Y} \mid \beta, \sigma^2 \sim N_n(\mathbf{X}\beta, \sigma^2 I), \tag{1}$$

where $\mathbf{Y}$ is $n \times 1$, $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$ is $n \times p$, $\beta = (\beta_1, \ldots, \beta_p)'$, and σ^2 is a scalar. Both β and σ^2 are considered unknown. For the model (1), selecting a subset of predictors is equivalent to setting to 0 those β_i's corresponding to the nonselected predictors.

We shall assume throughout that $\mathbf{X}_1, \ldots, \mathbf{X}_p$ contains no variable that would be included in every possible model. If this is not the case for some subset of $\mathbf{X}_1, \ldots, \mathbf{X}_p$, say $\mathbf{X}_1^*$,

* Edward I. George is Professor of Statistics, Department of MSIS, University of Texas, Austin, TX 78759-1175. Robert E. McCulloch is Associate Professor of Statistics, Graduate School of Business, University of Chicago, IL 60637. The authors thank Jay Kadane, Yum-Keung Kwan, Colin Mallows, John Tukey, and anonymous referees for very helpful remarks. This work was supported by the IBM Faculty Research Fund at the University of Chicago Graduate School of Business.

© 1993 American Statistical Association
Journal of the American Statistical Association
September 1993, Vol. 88, No. 423, Theory and Methods

This content downloaded from 98.249.87.104 on Sun, 03 Jun 2018 21:38:20 UTC
All use subject to http://about.jstor.org/terms

$\ldots, \mathbf{X}_r^*$, then $\mathbf{X}_1^*, \ldots, \mathbf{X}_r^*$ should be removed from $\mathbf{X}_1, \ldots, \mathbf{X}_p$, and $\mathbf{Y}$ and the remaining $\mathbf{X}_i$ should be replaced by the residual vectors $(\mathbf{I} - \mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime})\mathbf{Y}$ and $(\mathbf{I} - \mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime})\mathbf{X}_i$, $(\mathbf{X}^* \equiv [\mathbf{X}_1^*, \ldots, \mathbf{X}_r^*])$. This reduction can be motivated from a Bayesian perspective as initially integrating out the coefficients corresponding to $\mathbf{X}_1^*, \ldots, \mathbf{X}_r^*$ with respect to the uniform prior (Lebesgue measure). For example, if an intercept was to be included in every model (as is usually the case), then one should exclude $\mathbf{1}_p \equiv [1, \ldots, 1]$ from the set of potential predictors and replace $\mathbf{Y}$ and the $\mathbf{X}_i$ by their centered counterparts $(\mathbf{I} - \mathbf{1}_p\mathbf{1}_p'/n)\mathbf{Y}$ and $(\mathbf{I} - \mathbf{1}_p\mathbf{1}_p'/n)\mathbf{X}_i$. (Note that after such a transformation (1) no longer holds, because the components of $\mathbf{Y}$ would not be independent. As it turns out, this does not matter because the likelihood function of β is the same as if one had assumed independence.)

To extract information relevant to variable selection, we consider (1) as part of a larger hierarchical model. The key feature of this hierarchical model is that each component of β is modeled as having come from a mixture of two normal distributions with different variances. A similar setup in this context was considered by Mitchell and Beauchamp (1988), who instead used "spike and slab" mixtures. An important distinction of our approach is that we do not put a probability mass on $\beta_i = 0$.

By introducing the latent variable $\gamma_i = 0$ or 1, we represent our normal mixture by

$$\beta_i \mid \gamma_i \sim (1 - \gamma_i)N(0, \tau_i^2) + \gamma_i N(0, c_i^2\tau_i^2) \tag{2}$$

and

$$P(\gamma_i = 1) = 1 - P(\gamma_i = 0) = p_i. \tag{3}$$

As will be seen, the introduction of γ_i facilitates our analysis of the problem. Our use of it is based on the data augmentation idea of Tanner and Wong (1987). Diebolt and Robert (in press) have also successfully used this approach in the mixture context.

When $\gamma_i = 0$, $\beta_i \sim N(0, \tau_i^2)$, and when $\gamma_i = 1$, $\beta_i \sim N(0, c_i^2\tau_i^2)$. Our interpretation of this formulation is as follows. First, we set τ_i (>0) small so that if $\gamma_i = 0$, then β_i would probably be so small that it could be "safely" estimated by 0. Second, we set c_i large ($c_i > 1$ always) so that if $\gamma_i = 1$, then a non-0 estimate of β_i should probably be included in the final model. Specific choices of τ_i and c_i for this purpose are recommended in the next section. Based on this interpretation, p_i may be thought of as the prior probability that β_i will require a non-0 estimate, or equivalently that $\mathbf{X}_i$ should be included in the model.

To obtain (2) as the prior for $\beta_i \mid \gamma_i$, we use a multivariate normal prior

$$\beta \mid \gamma \sim N_p(0, \mathbf{D}_\gamma\mathbf{R}\mathbf{D}_\gamma), \tag{4}$$

where $\gamma = (\gamma_1, \ldots, \gamma_p)$, $\mathbf{R}$ is the prior correlation matrix, and

$$\mathbf{D}_\gamma \equiv \operatorname{diag}[a_1\tau_1, \ldots, a_p\tau_p], \tag{5}$$

with $a_i = 1$ if $\gamma_i = 0$ and $a_i = c_i$ if $\gamma_i = 1$. $\mathbf{D}_\gamma$ determines the scaling of the prior covariance matrix in such a way that (2) is satisfied. Here too, we set $\tau_1, \ldots, \tau_p$ small and $c_1, \ldots, c_p$ large ($c_i > 1$ always) so that under (4), those β_i for which $\gamma_i = 0$ will tend to be clustered around 0, whereas those β_i for which $\gamma_i = 1$ will tend to be dispersed. Recommended choices for the constants $\tau_1, \ldots, \tau_p$ and $c_1, \ldots, c_p$ and the prior correlation matrix $\mathbf{R}$ are discussed in the next section.

The Bernoulli model (3) is obtained as the marginal of any discrete distribution $f(\gamma)$ with support on the 2^p possible values of γ. But for the purpose of variable selection, $f(\gamma)$ should be the statistician's prior probability that γ correctly identifies (by $\gamma_i = 1$) exactly those β_i that should obtain non-0 estimates in the final model. Coupled with $f(\gamma)$, the prior on β is a finite mixture of multivariate normal priors.

The final ingredient in our hierarchical model is a prior on the residual variance σ^2. For this purpose we use the inverse gamma conjugate prior

$$\sigma^2 \mid \gamma \sim \mathrm{IG}(\nu_\gamma/2, \nu_\gamma\lambda_\gamma/2), \tag{6}$$

which is equivalent to $\nu_\gamma\lambda_\gamma/\sigma^2 \sim \chi^2_{\nu_\gamma}$. Note that ν_γ and λ_γ may depend on γ to incorporate dependence between β and σ^2. For example, if the regression were being used to approximate a complex relationship, then the user might expect that σ^2 would decrease as the dimension of β (= the number of non-0 components of γ) increased.

2. IDENTIFYING THE BEST MODELS WITH $f(\gamma \mid Y)$

Our main reason for embedding the normal linear model (1) in the hierarchical mixture model of Section 1 is to obtain the marginal posterior distribution $f(\gamma \mid \mathbf{Y}) \propto f(\mathbf{Y} \mid \gamma)f(\gamma)$, which contains the information relevant to variable selection. As described in Section 1, $f(\gamma)$ may be interpreted as the statistician's prior probability that the $\mathbf{X}_i$'s corresponding to non-0 components of γ (and only those $\mathbf{X}_i$'s), should be included in the final model. Based on the data $\mathbf{Y}$, the posterior $f(\gamma \mid \mathbf{Y})$ updates the prior probabilities on each of the 2^p possible values of γ. Identifying each γ with a submodel via $(\gamma_i = 1) \Leftrightarrow (\mathbf{X}_i \text{ is included})$, those γ with higher posterior probability $f(\gamma \mid \mathbf{Y})$ identify the submodels supported most by the data and the statistician's prior information. Thus $f(\gamma \mid \mathbf{Y})$ provides a ranking that can be used to select the more promising submodels for further investigation. We now proceed to discuss the choice of the prior $f(\gamma)$, the constants $\tau_1, \ldots, \tau_p$ and $c_1, \ldots, c_p$ for $\mathbf{D}_\gamma$ in (5), the prior correlation matrix $\mathbf{R}$ in (4), and ν_γ and λ_γ in (6). The ultimate value of the ranking provided by $f(\gamma \mid \mathbf{Y})$ depends on these choices.

2.1 Choosing $f(\gamma)$

The choice of $f(\gamma)$ should incorporate any available prior information about which subsets of $\mathbf{X}_1, \ldots, \mathbf{X}_p$ should be included in the final model. Although this may seem difficult with 2^p possible choices, especially with large p, symmetry considerations may simplify this task. For example, a reasonable choice might have the γ_i's independent with marginal distributions (3), so that

$$f(\gamma) = \prod p_i^{\gamma_i}(1 - p_i)^{(1-\gamma_i)}. \tag{7}$$

Although (7) implies that the inclusion of $\mathbf{X}_i$ is independent of the inclusion of $\mathbf{X}_j$ for all $i \neq j$, we found it to work well in various situations. The uniform or "indifference" prior

This content downloaded from 98.249.87.104 on Sun, 03 Jun 2018 21:38:20 UTC
All use subject to http://about.jstor.org/terms

$f(\gamma) \equiv 2^{-p}$ is the special case of (7) where each $\mathbf{X}_i$ has an equal chance ($p_i = \frac{1}{2}$) of being included. Alternatively, one may wish to weight more according to model size by using $f(\gamma) = w_{|\gamma|}\binom{p}{|\gamma|}^{-1}$, where $w_{|\gamma|}$ is the prior probability of a model of size $|\gamma|$. By setting $w_{|\gamma|}$ large for smaller $|\gamma|$, one can assign more weight to parsimonious models.

2.2 Choosing τ_i and c_i

The choice of τ_i in (2) and (5) should be such that if $\beta_i \sim N(0, \tau_i^2)$, then β_i can be "safely" replaced by 0. Because $|\beta_i| \leq 3\tau_i$ with high probability, as a rough guide the statistician may want to set $3\tau_i$ equal to the maximum size at which β_i would, for practical purposes, be equivalent to 0. Unfortunately, this may not be easy or even possible, because ascertaining this maximum requires understanding the potential effect of β_i in the final model. Thus alternative semiautomatic choices are discussed later in this section.

The choice of c_i (>1) in (2) and (5) should be such that if $\beta_i \sim N(0, c_i^2\tau_i^2)$, then a non-0 estimate of β_i should be included in the final model. From a subjectivist Bayesian standpoint, one would want to choose c_i large enough to give support to values of β_i that are substantively different from 0, but not so large that unrealistic values of β_i are supported. To help guide the choice of c_i, it may be useful to observe that the densities of $N(0, \tau_i^2)$ and $N(0, c_i^2\tau_i^2)$ intersect at $\xi(c_i)\tau_i$ when $\xi(c_i) = \sqrt{2(\log c_i)c_i^2/(c_i^2 - 1)}$. This implies that the density of $N(0, c_i^2\tau_i^2)$ will be larger than the density of $N(0, \tau_i^2)$ iff $|\beta_i| > \xi(c_i)\tau_i$. Note that this intersection point increases very slowly; for example, the choices $c_i = 10, 100, 1{,}000, 10{,}000, 100{,}000$ correspond to $\xi(c_i) \approx 2.1, 3.1, 3.7, 4.3, 4.8$. It may also be useful to observe that c_i is the ratio of the heights of $N(0, \tau_i^2)$ and $N(0, c_i^2\tau_i^2)$ at 0. Thus c_i can be interpreted as the prior odds that $\mathbf{X}_i$ should be excluded when β_i is very close to 0.

A semiautomatic approach to selecting τ_i and c_i may be obtained by considering the intersection point and relative heights at 0 of the marginal densities $(\hat{\beta}_i \mid \sigma_{\beta_i}, \gamma_i = 0) \sim N(0, \sigma_{\hat{\beta}_i}^2 + \tau_i^2)$ and $(\hat{\beta}_i \mid \sigma_{\beta_i}, \gamma_i = 1) \sim N(0, \sigma_{\hat{\beta}_i}^2 + c_i^2\tau_i^2)$. Let $t_i\sigma_{\beta_i}$ denote the intersection point, where $\sigma_{\hat{\beta}_i}^2$ is the variance of the least squares estimator $\hat{\beta}_i$. Because

$$P(\gamma_i = 1 \mid \hat{\beta}_i, \sigma_{\beta_i}) > p_i \,(= P(\gamma_i = 1)) \quad \text{iff } \hat{\beta}_i/\sigma_{\beta_i} > t_i, \tag{8}$$

the point t_i may be thought of as the threshold at which the t statistic corresponds to an increased marginal probability that $\mathbf{X}_i$ should be included in the model. Small t_i would tend to favor more saturated models, whereas large t_i would yield more parsimonious models. The relative heights of the marginal densities of $\hat{\beta}_i$ at 0 is easily seen to be

$$r_i \equiv \sqrt{\frac{\sigma_{\beta_i}^2/\tau_i^2 + c_i^2}{\sigma_{\beta_i}^2/\tau_i^2 + 1}}. \tag{9}$$

The value of r_i is the marginal posterior probability of including $\mathbf{X}_i$ when $\hat{\beta}_i = 0$. The reader should be cautioned that this univariate perspective may be slightly misleading, because our problem is fundamentally multivariate in nature. For example, model choice is indicated by $P(\gamma \mid \hat{\beta}, \sigma)$ rather than by the individual $P(\gamma_i \mid \hat{\beta}_i, \sigma_{\beta_i})$.

The values of t_i and r_i in (8) and (9) are functions only of σ_{β_i}/τ_i and c_i. Thus one might consider fixing $\hat{\sigma}_{\beta_i}/\tau_i$ and c_i to obtain desired value of t_i and r_i. ($\hat{\sigma}_{\beta_i}$ is the observed standard error commonly associated with the least squares estimate $\hat{\beta}_i$.) This approach also has the desirable feature of being invariant under rescaling of the $\mathbf{X}_i$'s. Some of the choices we consider in the examples are $(\sigma_{\beta_i}/\tau_i, c_i) = (1, 5), (1, 10), (10, 100), (10, 500)$, which yield $(t_i, r_i) \approx (2.4, 1.7), (2.7, 2.3), (2.1, 3.2), (2.8, 6.8)$. The marginal densities corresponding to these choices are displayed in Figure 1. Note that as σ_{β_i}/τ_i and c_i are increased, the separation between the two distributions becomes sharper.

In setting $\hat{\sigma}_{\beta_i}/\tau_i$, it is important to distinguish between the statistical significance of β_i, which is captured by $\hat{\sigma}_{\beta_i}$, and the practical significance of β_i, which is captured by τ_i. For example, if $\hat{\sigma}_{\beta_i}/\tau_i$ were set large, then β_i could be included in the model even when its sampling uncertainty overwhelmed its importance. On the other hand, if $\hat{\sigma}_{\beta_i}/\tau_i$ were set small, this would avoid including unimportant variables just because their effects were measured well.

It may be most productive to regard τ_i and c_i as tuning constants that calibrate the information in $f(\gamma \mid \mathbf{Y})$. Rather than treat any particular settings as hard and fast rules that guarantee good results, the user should consider varying these settings to extract more information. This strategy is illustrated on examples in Sections 4 and 5. Some caution might be exercised, however, because it follows from (2) and (3)

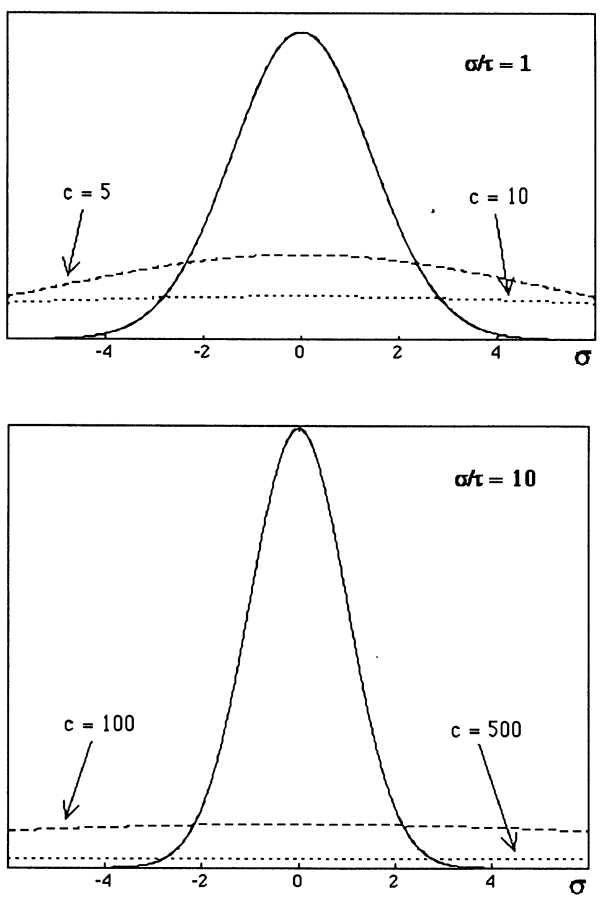

Figure 1. The Marginals $N(0, \sigma^2 + \tau^2)$ and $N(0, \sigma^2 + c^2\tau^2)(\sigma/\tau, c) = (1, 5), (1, 10), (10, 100), (10, 500)$.

This content downloaded from 98.249.87.104 on Sun, 03 Jun 2018 21:38:20 UTC
All use subject to http://about.jstor.org/terms

that for $0 < p_i < 1$ and $\beta_i \neq 0$, $\lim_{\tau_i \to 0} P(\gamma_i = 1 \mid \beta_i) = 1$ for fixed $c_i\tau_i$ and $\lim_{c_i \to \infty} P(\gamma_i = 1 \mid \beta_i) = 0$ for fixed τ_i.

2.3 Choosing R

The matrix $\mathbf{R}$ is the prior correlation matrix of β conditionally on γ. As in the case of choosing τ_i and c_i, we regard $\mathbf{R}$ as a tuning constant that calibrates the information in $f(\gamma \mid \mathbf{Y})$. In choosing $\mathbf{R}$, it may be useful to consider its effect on the posterior covariance matrix of β under $f(\beta \mid \mathbf{Y}, \sigma, \gamma)$, namely

$$(\sigma^{-2}\mathbf{X}'\mathbf{X} + \mathbf{D}_\gamma^{-1}\mathbf{R}^{-1}\mathbf{D}_\gamma^{-1})^{-1}. \tag{10}$$

Of particular interest may be the special cases $\mathbf{R} = \mathbf{I}$ and $\mathbf{R} \propto (\mathbf{X}'\mathbf{X})^{-1}$, which can be thought of as extremes. When $\mathbf{R} = \mathbf{I}$, the components of β are independent under $f(\beta \mid \gamma)$. When $\mathbf{R} \propto (\mathbf{X}'\mathbf{X})^{-1}$, the prior correlation is identical to the design correlation, a generalization of the g prior of Zellner (1986). From (10), one can see that under $\mathbf{R} = \mathbf{I}$ the posterior correlations will be less than those of the design correlation, whereas under $\mathbf{R} \propto (\mathbf{X}'\mathbf{X})^{-1}$ the posterior correlations will be identical to those of the design correlation. In cases of highly collinear regressors, one might also consider choices of $\mathbf{R}$ and $\tau_1, \ldots, \tau_p$ to reduce ill-conditioning of the posterior covariance (see Soofi 1990). Finally, one may want to consider putting a prior on $\mathbf{R}$, although this may greatly increase the computational requirements of our procedure.

2.4 Choosing ν_γ and λ_γ

In choosing ν_γ and λ_γ for the inverse gamma prior (6), one can make use of the interpretation that these carry information from an imaginary prior experiment where ν_γ is the number of observations and $[\nu_\gamma/(\nu_\gamma - 2)]\lambda_\gamma$ is the prior estimate of σ^2. Typically, these will be constant (i.e., $\nu_\gamma \equiv \nu$ and $\lambda_\gamma \equiv \lambda$ or will depend at most on γ only through $|\gamma|$, the number of non-0 components of γ. For example, one might let $[\nu_\gamma/(\nu_\gamma - 2)]\lambda_\gamma$ be a decreasing function of $|\gamma|$ when it is expected that higher-dimensional models will obtain a smaller σ^2. Finally, as will be seen from the posterior (14), the choice $\nu_\gamma \equiv 0$ (and any λ_γ) can be used to represent ignorance.

3. GIBBS SAMPLING THE BEST SUBSETS

As described in the preceding section, the first part of SSVS entails specifying the hierarchical normal mixture model so that the posterior $f(\gamma \mid \mathbf{Y})$ puts most weight on the more "promising" subsets of predictors. The second part, described here, entails extracting this information. Rather than calculate all 2^p posterior probabilities in $f(\gamma \mid \mathbf{Y})$, which would involve the same kind of computational burden we originally sought to avoid, SSVS uses the Gibbs sampler to generate a sequence

$$\gamma^1, \ldots, \gamma^m, \tag{11}$$

which in many cases converges rapidly in distribution to $\gamma \sim f(\gamma \mid \mathbf{Y})$. Such a sequence can be obtained quickly and efficiently, with far less effort than required to compute the entire posterior. Furthermore—and this is the crucial observation—the sequence in (11) will, with high probability in many cases, contain exactly the information relevant to variable selection. This is because those γ with highest probability will also appear most frequently and hence will be easiest to identify. Those γ that appear infrequently or not at all are simply not of interest and can be disregarded.

SSVS implements the Gibbs sampler to generate an auxiliary "Gibbs sequence"

$$\beta^0, \sigma^0, \gamma^0, \beta^1, \sigma^1, \gamma^1, \ldots, \beta^j, \sigma^j, \gamma^j, \ldots, \tag{12}$$

an ergodic Markov chain in which (11) is embedded. Except for β^0 and σ^0, which are initialized to be the least squares estimates of (1), and γ^0, which is initialized as $\gamma^0 \equiv (1, 1, \ldots, 1)$, the subsequent values of β^j, σ^j, γ^j are obtained by successively simulating values according to the following iterated sampling scheme. Fortunately, this scheme entails simulations that can be done fast and efficiently.

To begin, the coefficient vector β^j is obtained by sampling from

$$\beta^j \sim f(\beta^j \mid \mathbf{Y}, \sigma^{j-1}, \gamma^{j-1}) = N_p(\mathbf{A}_{\gamma^{j-1}}(\sigma^{j-1})^{-2}\mathbf{X}'\mathbf{X}\hat{\beta}_{\mathrm{LS}}, \mathbf{A}_{\gamma^{j-1}}), \tag{13}$$

where

$$\mathbf{A}_{\gamma^{j-1}} = ((\sigma^{j-1})^{-2}\mathbf{X}'\mathbf{X} + \mathbf{D}_{\gamma^{j-1}}^{-1}\mathbf{R}^{-1}\mathbf{D}_{\gamma^{j-1}}^{-1})^{-1}. \tag{13a}$$

Note that $\mathbf{D}_\gamma^{-1} = \mathrm{diag}[(a_1\tau_1)^{-1}, \ldots, (a_p\tau_p)^{-1}]$ is easily calculated. Next, the variance σ^j is obtained by sampling from

$$\sigma^j \sim f(\sigma^j \mid \mathbf{Y}, \beta^j, \gamma^{j-1}) = \mathrm{IG}\left(\frac{n + \nu_{\gamma^{j-1}}}{2}, \frac{|\mathbf{Y} - \mathbf{X}\beta^j|^2 + \nu_{\gamma^{j-1}}\lambda_{\gamma^{j-1}}}{2}\right), \tag{14}$$

the updated inverse gamma distribution from (6).

Finally, the vector γ^j is obtained componentwise by sampling consecutively (and preferably in random order) from the conditional distribution

$$\gamma_i^j \sim f(\gamma_i^j \mid \mathbf{Y}, \beta^j, \sigma^j, \gamma^j_{(i)}) = f(\gamma_i^j \mid \beta^j, \sigma^j, \gamma^j_{(i)}), \tag{15}$$

where $\gamma^j_{(i)} = (\gamma_1^j, \ldots, \gamma_{i-1}^j, \gamma_{i+1}^{j-1}, \ldots, \gamma_p^{j-1})$. Notice that the distribution (15) does not depend on $\mathbf{Y}$. This substantial simplification reduces computational requirements and allows for faster convergence of the subsequence (11). The nondependence of (15) on $\mathbf{Y}$ results from the hierarchical structure where γ affects $\mathbf{Y}$ only through β, a general feature of hierarchical models pointed out by Morris (1987).

Each distribution (15) is Bernoulli with probability

$$P(\gamma_i^j = 1 \mid \beta^j, \sigma^j, \gamma^j_{(i)}) = \frac{a}{a + b}, \tag{16}$$

where

$$a = f(\beta^j \mid \gamma^j_{(i)}, \gamma_i^j = 1) \times f(\sigma^j \mid \gamma^j_{(i)}, \gamma_i^j = 1) f(\gamma^j_{(i)}, \gamma_i^j = 1) \tag{16a}$$

and

$$b = f(\beta^j \mid \gamma^j_{(i)}, \gamma_i^j = 0) \times f(\sigma^j \mid \gamma^j_{(i)}, \gamma_i^j = 0) f(\gamma^j_{(i)}, \gamma_i^j = 0). \tag{16b}$$

It is worth noting that under the prior (7) on γ, and when the prior parameters for σ in (6) are constant ($\nu_\gamma \equiv \nu$ and λ_γ

This content downloaded from 98.249.87.104 on Sun, 03 Jun 2018 21:38:20 UTC
All use subject to http://about.jstor.org/terms

$\equiv \lambda$), (16) can be obtained more simply by

$$a = f(\boldsymbol{\beta}^j | \boldsymbol{\gamma}^j_{(i)}, \gamma^j_i = 1)p_i \tag{16c}$$

and

$$b = f(\boldsymbol{\beta}^j | \boldsymbol{\gamma}^j_{(i)}, \gamma^j_i = 0)(1 - p_i). \tag{16d}$$

Furthermore, under the choice of prior correlation $\mathbf{R} = \mathbf{I}$ in (4), the dependence on $\boldsymbol{\gamma}^j_{(i)}$ throughout (16) may be eliminated, further simplifying the calculations required.

By repeated successive sampling from (13), (14), and (15), the Gibbs sequence (12) is obtained. It follows from Diebolt and Robert (in press) that the subsequence (11) is a homogeneous ergodic Markov chain that converges geometrically to its unique stationary distribution $f(\boldsymbol{\gamma}|\mathbf{Y})$. A practical consequence of this property is that as the length of the subsequence (11) is increased, the empirical distribution of the realized values of $\boldsymbol{\gamma}$ will converge to the actual posterior $f(\boldsymbol{\gamma}|\mathbf{Y})$. Our experience has been that convergence appears to occur rapidly when $f(\boldsymbol{\gamma}|\mathbf{Y})$ is peaked, putting most of its mass on a few models. This is precisely when $f(\boldsymbol{\gamma}|\mathbf{Y})$ carries the most information about model selection.

At this point the information relevant to variable selection is contained in the sequence (11). In particular, after the sequence has reached approximate stationarity, the values of $\boldsymbol{\gamma}$ corresponding to the most promising subsets of $\mathbf{X}_1, \ldots, \mathbf{X}_p$ will appear with the highest frequency, because it is just those values which have largest probability under $f(\boldsymbol{\gamma}|\mathbf{Y})$. Thus a simple tabulation of the high-frequency values of $\boldsymbol{\gamma}$ can be used to identify the corresponding subsets of predictors as potentially promising. The potentially promising subsets of predictors may then be identified with these high-frequency values of $\boldsymbol{\gamma}$. The low-frequency or zero-frequency values of $\boldsymbol{\gamma}$ may simply be ignored, because these correspond to the least promising models. Note that if no high-frequency values of $\boldsymbol{\gamma}$ appeared in (11), then we would conclude that either m is too small or the data contain little information for discriminating between models.

It may also be fruitful to go beyond a simple tabulation of the high-frequency $\boldsymbol{\gamma}$ values in the sequence (11). For example, it is tempting to consider the marginal frequency of $\gamma_i = 1$ as evidence for the inclusion of $\mathbf{X}_i$. Unfortunately, this simplifying approach can be misleading unless there is little or no correlation among $\mathbf{X}_1, \ldots, \mathbf{X}_p$. In general, inclusion of $\mathbf{X}_i$ must be considered jointly with other variables, and so it may be better to look at conditional frequencies. Other approaches might include the exploratory data methods discussed in Tukey (1977), such as product-ratio plots of the $\boldsymbol{\gamma}$ frequency counts. Finally, one might consider a second iteration of SSVS with a reduced set of variables based on the first run.

Note that although there is dependence on initial values in the sequence, we do not recommend ignoring the first few values as is often done in applications of the Gibbs sampler. Essentially, we regard SSVS as exploratory and just want to ascertain which $\boldsymbol{\gamma}$ may have large $f(\boldsymbol{\gamma}|\mathbf{Y})$. We do recommend that m be chosen as large as is economically feasible to mitigate any bias in the frequency estimates due to dependence in (11).

Finally, it should be mentioned that convergence of (11) can be very slow in certain situations. For example, this may occur when τ_i is chosen very small and c_i is chosen very large so that the prior for β_i is close to a "spike and slab" mixture. Apparently, this setup can lead to very small transition probabilities for γ^j_i in the Gibbs sequence to go from 0 to 1 or from 1 to 0. This phenomenon can be further exacerbated by the presence of more than one model with high probability under $f(\boldsymbol{\gamma}|\mathbf{Y})$ with very small transition probabilities between models. In these cases the Gibbs sequence frequencies may take a long time to converge to $f(\boldsymbol{\gamma}|\mathbf{Y})$. To avoid such problems, we recommend experimentation with various values for τ_i and c_i.

4. SIMULATED EXAMPLES

In this section we illustrate the performance of SSVS on simulated examples. Example 4.1 treats small problems involving five potential predictors. Example 4.2 considers a large problem with 60 potential predictors, which is currently about double the maximum size at which conventional all-subsets regression strategies can be carried out (see Miller 1990). This example demonstrates that SSVS is a feasible alternative that performs sensibly in such large problems.

Example 4.1. This example considers two simple, variable selection problems with $p = 5$ predictors of length $n = 60$. In Problem 1, the predictors were obtained as independent standard normal vectors, $\mathbf{X}_1, \ldots, \mathbf{X}_5$ iid $\sim N_{60}(0, 1)$, so that they were practically uncorrelated. The dependent variable was generated according to the model

$$\mathbf{Y} = \mathbf{X}_4 + 1.2\mathbf{X}_5 + \varepsilon, \tag{17}$$

where $\varepsilon \sim N_{60}(0, \sigma^2 I)$ with $\sigma = 2.5$. Thus $\boldsymbol{\beta} = (0, 0, 0, 1, 1.2)'$. The least squares estimates for these data were $\hat{\boldsymbol{\beta}} = (.03, -.45, .23, .84, 1.29)'$, with standard errors $\hat{\sigma}_{\boldsymbol{\beta}} = (.36, .40, .36, .31, .33)'$ and $\hat{\sigma} = 2.58$.

Problem 2 is identical to Problem 1, except that $\mathbf{X}_3$ is replaced by $\mathbf{X}^*_3 = \mathbf{X}_5 + .15\mathbf{Z}$ where $\mathbf{Z} \sim N_{60}(0, 1)$, yielding $\text{corr}(\mathbf{X}_3, \mathbf{X}_5) = .989$. This $\mathbf{X}^*_3$ is a substantial proxy for $\mathbf{X}_5$. Problem 2 is meant to illustrate how SSVS performs in the presence of extreme collinearity. The least squares estimates for these data were $\hat{\boldsymbol{\beta}} = (.01, -.38, .34, .83, .95)'$, with standard errors $\hat{\sigma}_{\boldsymbol{\beta}} = (.35, .39, 2.33, .31, 2.35)'$ and $\hat{\sigma} = 2.59$. Although the coefficient estimates are nearly the same as those obtained in Problem 1, the standard errors for $\hat{\beta}_3$ and $\hat{\beta}_5$ are substantially increased due to the induced collinearity.

We applied SSVS to both problems with the indifference prior $f(\boldsymbol{\gamma}) \equiv \frac{1}{2}^5$, $\tau_1 = \cdots = \tau_5 = .33$, $c_1 = \cdots = c_5 = 10$, $\mathbf{R} = \mathbf{I}$, and $\nu_{\boldsymbol{\gamma}} \equiv 0$. This setting yields $\hat{\sigma}_{\beta_i}/\tau_i \approx 1$, which, coupled with $c_i = 10$, corresponds to one of the settings in Figure 1. For this setting the threshold for inclusion in the model occurs at coefficient estimates with a t statistic of about 2.7. A sample of $m = 5{,}000$ observations of the Gibbs sequence (11) was then simulated and tabulated. Table 1 displays the four highest-frequency values of $\boldsymbol{\gamma}$ that appeared in each problem.

For Problem 1, the two most frequent models—$\hat{\mathbf{Y}} = f(\mathbf{X}_5)$ and $\hat{\mathbf{Y}} = f(\mathbf{X}_4, \mathbf{X}_5)$—appeared with frequencies of 25.8% and 24.2%. Although $\beta_4 = 1$ is non-0, $\mathbf{X}_4$ was often excluded

This content downloaded from 98.249.87.104 on Sun, 03 Jun 2018 21:38:20 UTC
All use subject to http://about.jstor.org/terms

Table 1. High Frequency Models, Example 4.1

Problem 1		Problem 2	
Model variables	Proportion	Model variables	Proportion
5	.258	3	.146
4 5	.242	5	.123
2 5	.070	3 4	.098
2 4 5	.055	4 5	.086

because the t statistic for $\hat{\beta}_4$ was $.84/.31 \approx 2.7$, just about equal to the inclusion threshold for $\hat{\sigma}_{\beta_i}/\tau_i \approx 1$ and $c_i = 10$. This illustrates how SSVS is set up to exclude variables whose coefficients are "close" to 0 relative to this threshold. For the other variables, only the t statistic for $\hat{\beta}_5$—$1.29/.33 \approx 3.9$—was larger than this threshold. The other two Problem 1 models in Table 1, which occurred less frequently, also included $\mathbf{X}_4$ and/or $\mathbf{X}_5$ but sometimes allowed $\mathbf{X}_2$ to stray in. This shows how SSVS is useful in identifying several promising models rather than the single best model. This feature is similar to the way in which stepwise methods are used to narrow the scope of model selection.

For Problem 2, each model containing $\mathbf{X}_5$ occurred with nearly the same frequency as the corresponding model with $\mathbf{X}_5$ replaced by $\mathbf{X}_3$. Furthermore, if $\mathbf{X}_3$ and $\mathbf{X}_5$ are considered identical (which they nearly are), then these output frequencies are essentially the same as those for Problem 1. Of course when one has strong proxies, either one will do, so that our procedure is still identifying the more promising models. Unfortunately, this example illustrates how introducing proxies may dilute the focus of SSVS by increasing the number of promising models. To avoid this dilution, it may be worthwhile to eliminate strong proxies from the data before using SSVS. Finally, note that this example shows how marginal frequencies by themselves do not tell the whole story. Although $\mathbf{X}_5$ is equally effective in both problems, it appears in fewer models here because of the proxy $\mathbf{X}_3$.

Example 4.2. This example is meant to demonstrate the practical potential of SSVS for data sets involving many potential predictors. We constructed $p = 60$ predictors, $\mathbf{X}_1, \ldots, \mathbf{X}_{60}$, of length $n = 120$. These were obtained as $\mathbf{X}_i = \mathbf{X}_i^* + \mathbf{Z}$, where $\mathbf{X}_1^*, \ldots, \mathbf{X}_{60}^*$ iid $\sim N_{120}(0, 1)$ independently of $\mathbf{Z} \sim N_{120}(0, 1)$. This induced pairwise correlations of about .5. The dependent variable was generated according to the model $\mathbf{Y} = [\mathbf{X}_1, \ldots, \mathbf{X}_{60}]\beta + \varepsilon$, where $\varepsilon \sim N_{120}(0, \sigma^2 I)$ with $\sigma = 2$, and the coefficients $\beta = (\beta_1, \ldots, \beta_{60})'$ were set at $(\beta_1, \ldots, \beta_{15}) = (0, \ldots, 0)$, $(\beta_{16}, \ldots, \beta_{30}) = (1, \ldots, 1)$, $(\beta_{31}, \ldots, \beta_{45}) = (2, \ldots, 2)$, and $(\beta_{46}, \ldots, \beta_{60}) = (3, \ldots, 3)$.

We applied SSVS with the indifference prior $f(\gamma) \equiv \frac{1}{2}^{60}$, $\mathbf{R} = \mathbf{I}$, $\nu_\gamma \equiv 0$, and the four automatic settings $(\hat{\sigma}_{\beta_i}/\tau_i, c_i) = (1, 5), (1, 10), (10, 100), (10, 500)$ discussed in Section 2.2. For each setting, a sample of $m = 30{,}000$ observations of the Gibbs sequence (11) was then simulated and tabulated. It took 31 seconds per 1,000 iterations, or 15.5 minutes total, to generate each sample using Fortran compiled with the fast option on a Sun Sparcstation 10. Table 2 lists the five highest-frequency models and Figure 2 displays all the frequency counts for each of the four simulated sequences.

Table 2 shows that under the two settings (1, 5) and (10, 100), SSVS is doing extremely well. The highest-frequency model under both settings was the "correct" model, and the next four most frequent models had only one variable (in one case, two variables) "incorrectly" included or excluded. The other two settings (1, 10) and (10, 500) did only slightly worse, incorrectly excluding variables with small non-0 coefficients. Apparently, this resulted from increasing the c_i's, and thereby lowering the probabilities of inclusion.

As shown in Figure 2, the distribution of γ frequencies for each setting is extremely J-shaped, with very few models appearing with high frequency. Furthermore, as $(\hat{\sigma}_{\beta_i}/\tau_i, c_i)$ is varied along the four settings, fewer models are visited. Indeed, out of 2^{60} possible models under (1, 5), (1, 10), (10, 100), (10, 500), the total number of models visited was 25,723, 22,468, 5,071 and 1,901. Of these, 23,384, 19,390, 2,171 and 511 were visited only once. As shown in Table 2, the highest frequencies increase dramatically along these settings. It appears that as $(\hat{\sigma}_{\beta_i}/\tau_i, c_i)$ is increased, the posterior distribution $f(\gamma|\mathbf{Y})$ becomes more peaked. This suggests that the statistician may be able to vary $(\hat{\sigma}_{\beta_i}/\tau_i, c_i)$ to focus this posterior on a small set of models. This underscores our suggestion to vary τ_i and c_i to identify promising models.

5. REAL DATA EXAMPLES

In this section we apply SSVS to three real examples. The first Example 5.1 illustrates the performance of SSVS on a familiar data set. Example 5.2 illustrates how prior information can be usefully incorporated into the hierarchical setup that is the basis for SSVS. Example 5.3 illustrates the performance of SSVS on a large problem.

Table 2. High Frequency Models, Example 4.2

(1, 5)		(1, 10)		(10, 100)		(10, 500)	
Freq	False choice	Freq	False choice	Freq	False choice	Freq	False choice
51	—	57	16, 20, 26, 30	2457	—	2845	17, 20, 26, 30
28	3	53	17, 20, 26, 30	770	30	2109	17, 20, 22, 26, 30
26	30	53	16, 30	755	14	1407	17, 19, 20, 22, 26, 30
25	14	52	30	411	16, 30	761	17, 19, 20, 26, 30
21	11	49	16, 26, 30	406	3	745	16, 20, 26, 30

NOTE: False choice is false inclusion for variables 1–15, and false exclusion for variables 16–60.

This content downloaded from 98.249.87.104 on Sun, 03 Jun 2018 21:38:20 UTC
All use subject to http://about.jstor.org/terms

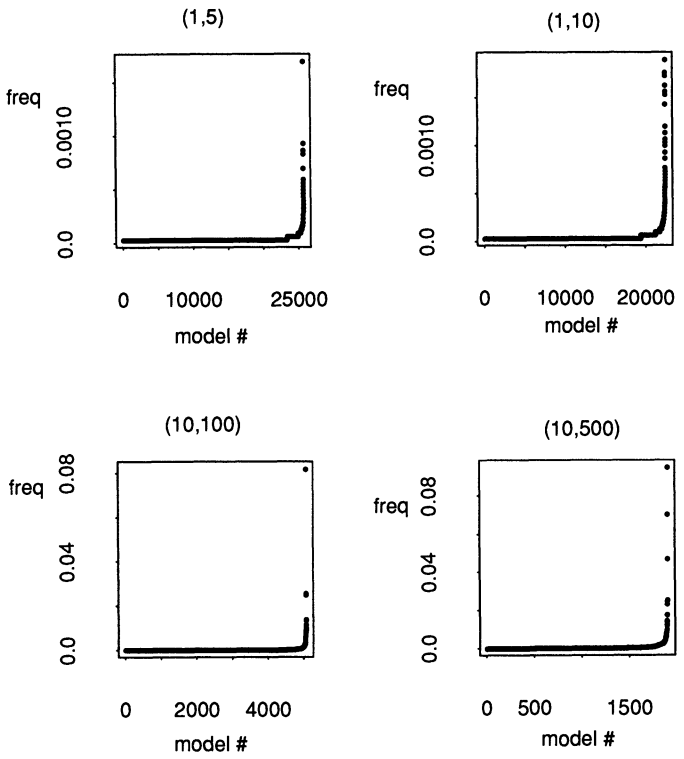

Figure 2. Model Frequencies, Example 4.2.

Example 5.1. The data for our first real example is the familiar Hald data (Draper and Smith 1981), which have been used by various authors to illustrate variable selection procedures. The data consist of $n = 13$ observations on a dependent variable Y (heat evolved during a chemical reaction) and $p = 4$ independent variables X_1, X_2, X_3, X_4 (inputs to the reaction). Thus $2^4 = 16$ possible models are under consideration. (An intercept is always included.) As described by Draper and Smith (1981), three models were favored by conventional selection procedures. The model $\hat{Y} = f(X_1, X_2)$, yielding $R^2 = 97.9\%$, was favored by all subsets regression, backward elimination, and stepwise regression; the model $\hat{Y} = f(X_1, X_4)$, yielding $R^2 = 97.2\%$, was also favored by all subsets regression; and the model $\hat{Y} = f(X_1, X_2, X_4)$, yielding $R^2 = 98.2\%$, was favored by forward selection.

For the purpose of comparison, we applied SSVS to the Hald data with the indifference prior $f(\gamma) \equiv \frac{1}{2}^4$, $\mathbf{R} = \mathbf{I}$, $\nu_\gamma \equiv 0$, and the four automatic settings $(\hat{\sigma}_{\beta_i}/\tau_i, c_i) \equiv (1, 5), (1, 10), (10, 100), (10, 500)$ discussed in Section 2.2. For each setting, a sample of $m = 5{,}000$ observations of the Gibbs sequence [11] was then simulated and tabulated. Table 3 lists the model frequencies for all γ's that realized a frequency of at least 2% for some sequence. Under (1, 5) and (1, 10), SSVS puts more probability on models with few or no variables. In contrast, under (10, 100) and (10, 500), SSVS puts more probability on models chosen by conventional procedures. Furthermore, under these latter settings, the posterior distribution is much more concentrated around a smaller number of models, just as in Example 4.2.

Example 5.2. The data for our second real example were collected to test the hypothesis that "love" and "work" are the important factors in determining an individual's happiness. As alternatives, the variables "money" and "sex" were included in the study. (Here "sex" refers to sexual activity rather than gender.) Five variables were recorded: Y = Happiness, X_1 = Money, X_2 = Sex, X_3 = Love, and X_4 = Work. Happiness was measured on a 10-point scale, with 1 representing a suicidal state, 5 representing a feeling of "just muddling along," and 10 representing a euphoric state. Money was measured by annual family income in thousands of dollars. Sex was measured by a dummy variable taking the values 0 or 1, with 1 indicating a satisfactory level of sexual activity. Love was measured on a 3-point scale, with 1 representing loneliness and isolation, 2 representing a set of secure relationships, and 3 representing a deep feeling of belonging and caring in the context of some family or community. Work was measured on a 5-point scale, with 1 indicating that an individual is seeking other employment, 3 indicating the job is "OK," and 5 indicating that the job is enjoyable. The data were collected from the 39 individuals in an MBA class for employed students at the University of Chicago Graduate School of Business.

To implement the SSVS procedure in this example, it was possible to base the choice of τ_i on practical considerations. To begin, although the data exhibit correlation among the four explanatory variables, we were interested in each effect while holding the other variables constant. The maximum effect of a variable X_i was then considered to be $\beta_i \Delta X_i$, where ΔX_i is the maximum change we would expect in X_i. Letting ΔY be the threshold at which the effect is unimportant, we then set $\tau_i = \Delta Y / 3\Delta X_i$, because $\Delta Y / \Delta X_i$ would be the corresponding threshold for each β_i. For example, because X_2 is a dummy variable, $\Delta X_2 = 1$ and $\tau_2 = \Delta Y / 3$.

For the purpose of robustness, we considered various choices of τ_i and c_i. For each τ_i we considered the "low" and "high" settings, $\tau_i = .5/3\Delta X_i$ and $\tau_i = 1/3\Delta X_i$, cor-

Table 3. Model Frequencies for the Hald Data, Example 5.1

Model variables	(1, 5)	(1, 10)	(10, 100)	(10, 500)
NONE	.23	.44	—	—
1	.26	.25	—	—
2	.06	.05	—	—
3	.06	.06	—	—
4	.08	.07	—	—
1, 2	.07	.03	.60	.81
1, 3	.06	.03	—	—
1, 4	.06	.03	.27	.16
1, 2, 3	.02	.00	.03	.01
1, 2, 4	.02	.00	.05	.01
1, 3, 4	.01	.00	.05	.01

NOTE: Models visited less than 2% for all settings not listed. "—" indicates model not visited.

Table 4. The Six Priors, Example 5.2

π	1	2	3	4	5	6
τ_i	$.5/3\Delta X_i$	$.5/3\Delta X_i$	$1/3\Delta X_i$	$1/3\Delta X_i$	$1/3\Delta X_i$	$\hat{\sigma}_{\beta_i}/10$
c_i	4	9	4	9	4	100
R	I	I	I	I	$\propto (\mathbf{X}'\mathbf{X})^{-1}$	I

This content downloaded from 98.249.87.104 on Sun, 03 Jun 2018 21:38:20 UTC
All use subject to http://about.jstor.org/terms

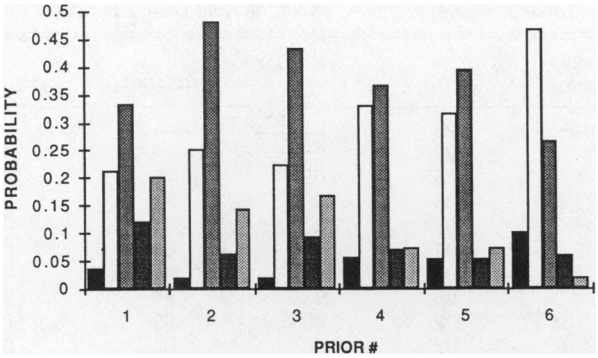

Figure 3. Model Probabilities, Example 5.2. ■, X3; □, X3, X4; ■, X1, X3, X4; ■, X2, X3, X4; ▩, X1, X2, X3, X4.

responding to the threshold choices $\Delta Y = .5$ and $\Delta Y = 1$. For each c_i we considered the "low" and "high" settings, $c_i = 4$ and $c_i = 9$. These choices provided substantial separation between the two mixture components in (2) while still allowing for plausible values of β_i when $\gamma_i = 1$. Because we favored no particular γ, we used the indifference prior $f(\gamma) \equiv \frac{1}{2}^4$. (An intercept was always included.) Finally, we chose $\nu_\gamma \equiv 3$ and $\lambda_\gamma \equiv 25$. With these choices, the expected value of σ is about 7 and the probability that σ is between 2.6 and 26.5 is about .98.

After centering the data to account for the intercept, we applied SSVS to this data using the six prior choices listed in Table 4, with $f(\gamma) \equiv \frac{1}{2}^4$, $\nu_\gamma \equiv 3$, and $\lambda_\gamma \equiv 25$. Note that priors 1–4 vary the settings for all τ_i and all c_i between "low" and "high." Prior 5 takes our preferred settings for τ_i and c_i and uses the g prior $\mathbf{R} \propto (\mathbf{X}'\mathbf{X})^{-1}$, and prior 6 uses the default $\hat{\sigma}_{\beta_i}/\tau_i = 10$ with $c_i = 100$.

Figure 3 displays the five high-frequency models of each size obtained by each of the six priors. For every prior the two most probable models selected were (3 4) and (1 3 4). Prior 1, which had smallest τ_i and c_i, seemed to favor more saturated models and gave more weight to (1 2 3 4) than did the other priors. In contrast, prior 6, which used the default setting with large c_i, seemed to favor more parsimonious models and was the only prior to put more weight on (3 4) than on (1 3 4). Aside from these differences, there seemed to be substantial agreement of model frequencies for the six priors, suggesting reasonable robustness of SSVS with respect to prior specification. Figure 4 displays the marginal probability estimates for each of the four variables with the six priors. Every prior ordered the variables in importance as X_3, X_4, X_1, X_2. Here the agreement of frequencies for the different priors is striking.

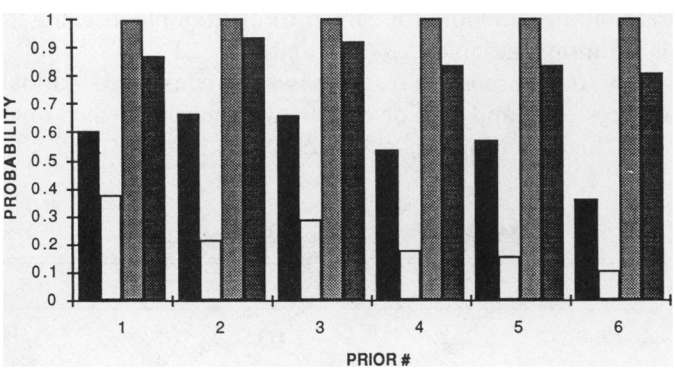

Figure 4. Marginal Probabilities, Example 5.2. ■, X1; □, X2; ■, X3; ■, X4.

Example 5.3. The third real data set we considered was collected by a bank to study the selling of new accounts and to characterize each branch office in this regard. The data consisted of 233 observations on each of the following 16 variables: Y = the number of new accounts sold in a given time period, X_1 = number of households serviced, X_2 = number of people selling the new account, $X_3 = 1$ if the branch is in Manhattan and 0 otherwise, $X_4 = 1$ if the branch is in the boroughs and 0 otherwise, $X_5 = 1$ if the branch is in the suburbs and 0 otherwise, X_6 = demand deposits balance, X_7 = number of demand deposit, X_8 = now accounts balance, X_9 = number of now accounts, X_{10} = balance of money market accounts, X_{11} = number of money market accounts, X_{12} = passbook saving balance, X_{13} = other time balance, X_{14} = consumer loans, and X_{15} = shelter loans.

Table 5 is the summary output for a least squares regression on all 15 variables (including an intercept). Residual plots supported the usual normal linear model assumptions. In this full model, the weakest variables X_4, X_5, and X_{12} obtained p values larger than .5. Using forward selection, the last five variables to be entered were $X_4, X_{11}, X_{10}, X_{12}$, and X_5. These were also the first five removed by backward elimination, in reverse order. A minimum C_p value of 12.3 and $R^2 = 91.4\%$ was achieved by the model that excluded only X_5 and X_{12}.

After centering the data to account for a constant term, SVSS was run on this data with the indifference prior $f(\gamma) \equiv \frac{1}{2}^{15}$, the setting $(\hat{\sigma}_{\beta_i}/\tau_i, c_i) = (10, 100)$ discussed in Section 2.2, $\nu_\gamma \equiv 0$, and the two prior covariance choices $\mathbf{R} = \mathbf{I}$ and $\mathbf{R} \propto (\mathbf{X}'\mathbf{X})^{-1}$. For each of these two priors, a Gibbs sequence (11) of length $m = 10{,}000$ was simulated.

Both simulated sequences yielded J-shaped frequency distributions of γ values similar to those in Figure 2. Totals of 696 and 1,278 distinct γ values were visited under $\mathbf{R} = \mathbf{I}$ and $\mathbf{R} \propto (\mathbf{X}'\mathbf{X})^{-1}$ respectively. Of these only 19 and 14 γ

Table 5. The Full Least Squares Regression, Example 5.3

Predictor	Coef	Stdev	t-ratio	P
Constant	19.62	23.22	0.84	0.399
X_1	−0.015298	0.004949	−3.09	0.002
X_2	11.476	4.056	2.83	0.005
X_3	48.22	23.74	2.03	0.043
X_4	15.38	22.87	0.67	0.502
X_5	−8.40	21.62	−0.39	0.698
X_6	−0.019851	0.004308	−4.61	0.000
X_7	0.025384	0.009497	2.67	0.008
X_8	−0.013079	0.004366	−3.00	0.003
X_9	0.54401	0.05242	10.38	0.000
X_{10}	−0.002289	0.001568	−1.46	0.146
X_{11}	0.09234	0.04367	2.11	0.036
X_{12}	0.000976	0.002480	0.39	0.694
X_{13}	0.005675	0.002203	2.58	0.011
X_{14}	0.36340	0.04419	8.22	0.000
X_{15}	0.008022	0.002572	3.12	0.002

$s = 49.91$ $R - sq = 91.5\%$ $R - sq(adj) = 90.9\%$

This content downloaded from 98.249.87.104 on Sun, 03 Jun 2018 21:38:20 UTC
All use subject to http://about.jstor.org/terms

Table 6. High Frequency Models for (10, 100), Example 5.3

$R = I$					$R = (X'X)^{-1}$				
Freq	Exclusion	R^2	p	Cp	Freq	Exclusion	R^2	p	Cp
827	4, 5	91.2	14	18.4	1452	—	91.5	16	16.0
627	2, 4, 5	90.8	13	27.6	782	3	91.3	15	18.1
595	3–5, 11	90.4	12	34.8	718	4, 5	91.2	14	18.4
486	3–5	90.5	12	34.8	633	4	91.4	15	14.4
456	3, 4	91.0	13	22.8	564	5	91.5	15	14.1
390	4, 5, 11	91.1	13	19.5	493	3, 4	91.0	14	22.8
315	2–4	90.6	13	32.3	379	2	91.1	15	22.0
245	3, 4, 11	90.9	13	24.7	346	3–5	90.5	13	34.8
209	2, 4, 5, 11	90.6	12	29.2	249	2, 4	91.1	14	21.0
209	2, 4	91.1	14	21.0	193	2, 5	91.1	14	20.0

values were visited more than 100 times under $\mathbf{R} = \mathbf{I}$ and $\mathbf{R} \propto (\mathbf{X}'\mathbf{X})^{-1}$. Table 6, which lists the five highest-frequency models appearing in each sequence, shows that many different models of different sizes were selected by SSVS. Furthermore, every model in Table 6 is different from all the models obtained by the stepwise procedures. The minimum C_p model did appear in the Gibbs sequences, although never with a frequency higher than 10. Interestingly, the models in Table 6 excluded X_4, X_5, or X_{11} but not X_{10} or X_{12}. All of the models in Table 6 obtained R^2 values practically as large as the full model in Table 5. It was also interesting that although there was some overlap in the models selected using $\mathbf{R} = \mathbf{I}$ and using $\mathbf{R} \propto (\mathbf{X}'\mathbf{X})^{-1}$, there was a pronounced difference. Using $\mathbf{R} = \mathbf{I}$, which tends to lessen posterior correlations, tended to select smaller models than using $\mathbf{R} \propto (\mathbf{X}'\mathbf{X})^{-1}$, which tends to replicate the design correlation. Certain variables, such as X_{11}, were excluded more often under $\mathbf{R} = \mathbf{I}$ than under $\mathbf{R} \propto (\mathbf{X}'\mathbf{X})^{-1}$.

Finally, this example once again illustrates how SSVS narrows the scope of possible models for further consideration. The choice of a single "best" model at this point could proceed by applying standard model selection criteria, such as C_p plots or predictive error measures, to the more manageable selected subset. Another approach might be to average the selected models (see Madigan and Raftery 1991).

[*Received February 1991. Revised October 1992.*]

REFERENCES

Casella, G., and George, E. I. (1992), "Explaining the Gibbs Sampler," *The American Statistician,* 46, 167–174.

Diebolt, J., and Robert, C. (in press), "Estimation of Finite Mixture Distributions Through Bayesian Sampling, *Journal of the Royal Statistical Society,* Ser. B.

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

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling," *Journal of the American Statistical Association,* 85, 972–985.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association,* 85, 398–409.

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

Madigan, D., and Raftery, A. E., (1991), "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," Technical Report 213, University of Washington, Dept. of Statistics.

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

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.

Morris, C. N., (1987), Comment on "The Calculation of Posterior Distributions by Data Augmentation" by M. A. Tanner and W. H. Wong, *Journal of the American Statistical Association,* 82, 542–543.

Pericchi, L. R. (1984), "An Alternative to the Standard Bayesian Procedure for Discrimination Between Normal Linear Models," *Biometrika,* 71, 575–586.

Poirier, D. J. (1985), "Bayesian Hypothesis Testing in Linear Models With Continuously Induced Conjugate Priors Across Hypotheses," in *Bayesian Statistics 2,* eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, New York: Elsevier, pp. 711–722.

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.

Soofi, E. S. (1990), "Effects of Collinearity on Information About Regression Coefficients," *Journal of Econometrics,* 43, 255–274.

Spiegelhalter, D. J., and Smith, A. F. M. (1982), "Bayes Factors for Linear and Log-Linear Models With Vague Prior Information," *Journal of the Royal Statistical Society,* Ser. B, 44, 377–87.

Stewart, L. (1987), "Hierarchical Bayesian Analysis Using Monte Carlo Integration: Computing Posterior Distributions When There are Many Possible Models," *The Statistician,* 36, 211–219.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association,* 82, 528–550.

Tukey, J. W. (1977), *Exploratory Data Analysis,* Reading, MA: Addison-Wesley.

Verdinelli, I., and Wasserman, L. (1991), "Bayesian Analysis of Outlier Problems Using the Gibbs Sampler," *Statistics and Computing,* 1, 105–117.

Zellner, A. (1984), "Posterior Odds Ratios for Regression Hypotheses: General Considerations and Some Specific Results," in *Basic Issues in Econometrics,* ed. A. Zellner, Chicago: University of Chicago Press, pp. 275–305.

—— (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis with *g* Prior Distributions," in *Bayesian Inference and Decision Techniques,* eds. P. Goel and A. Zellner, New York: Elsevier, pp. 233–243.

This content downloaded from 98.249.87.104 on Sun, 03 Jun 2018 21:38:20 UTC
All use subject to http://about.jstor.org/terms