Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Shotgun Stochastic Search for "Large *p*" Regression

Chris Hans, Adrian Dobra & Mike West

To cite this article: Chris Hans, Adrian Dobra & Mike West (2007) Shotgun Stochastic Search for "Large *p*" Regression, Journal of the American Statistical Association, 102:478, 507-516, DOI: 10.1198/016214507000000121

To link to this article: https://doi.org/10.1198/016214507000000121

Published online: 01 Jan 2012.

Shotgun Stochastic Search for "Large *p*" Regression

Chris HANS, Adrian DOBRA, and Mike WEST

Model search in regression with very large numbers of candidate predictors raises challenges for both model specification and computation, for which standard approaches such as Markov chain Monte Carlo (MCMC) methods are often infeasible or ineffective. We describe a novel *shotgun stochastic search* (SSS) approach that explores "interesting" regions of the resulting high-dimensional model spaces and quickly identifies regions of high posterior probability over models. We describe algorithmic and modeling aspects, priors over the model space that induce sparsity and parsimony over and above the traditional dimension penalization implicit in Bayesian and likelihood analyses, and parallel computation using cluster computers. We discuss an example from gene expression cancer genomics, comparisons with MCMC and other methods, and theoretical and simulation-based aspects of performance characteristics in large-scale regression model searches. We also provide software implementing the methods.

KEY WORDS: Model averaging; Parallel computing; Regression model uncertainty; Stochastic search; Variable selection.

1. INTRODUCTION

Regression variable uncertainty—framed as either model selection or model averaging—raises modeling and computational challenges as the number of candidate predictor variables increases. Standard methods, including stepwise methods, best-subset regression (e.g., Furnival and Wilson 1974), and Markov chain Monte Carlo (MCMC), often can quickly identify "good" models when the number of predictors is relatively small. In higher-dimensional problems, stepwise methods are prone to entrapment in local maxima of model space (Hocking 1976), and often do not provide an adequate representation of the model space with the increasingly complex patterns of collinearity that are typical with many variables. MCMC algorithms designed to explore the posterior distribution over regression model spaces (e.g., George and McCulloch 1993, 1997; Green 1995; Madigan and York 1995; Geweke 1996; Raftery, Madigan, and Hoeting 1997; Brown, Vannucci, and Fearn 1998b) rely on Gibbs sampling (Gelfand and Smith 1990) or on the Metropolis–Hastings algorithm but are increasingly ineffective in higher dimensions due to slow convergence. Outside of the regression model context, MCMC approaches have been used for model space exploration by Chipman, George, and McCulloch (1998) for Bayesian CART models, Wong, Carter, and Kohn (2003) for covariance selection models, and Tadesse, Sha, and Vannucci (2005) for clustering.

Here we introduce a novel *shotgun stochastic search* (SSS) method that is inspired by MCMC but offers the ability to identify probable models much more rapidly and to move around swiftly in the space of models as the dimension escalates. Parallel computing is at the core of SSS methodology; rather than naively parallelizing existing MCMC stochastic search methods by running parallel chains, we describe a new stochastic search that is inspired by and related to Metropolis–Hastings methods but differs fundamentally in two key respects:

- SSS explores the vast discrete space of regression models by evaluating and recording many candidate models in parallel at each iteration; this contrasts with traditional MCMC methods, which move sequentially from one model to a new model and so do not exploit the opportunity to effectively explore the model space in the neighborhood of known "good" models.
- SSS is designed to move toward and aggressively explore *regions* of regression model space that contain multiple higher-probability models, because it looks at many neighbors of each model selected. Thus, in contrast to MCMC methods that move only toward individual models of higher probability, SSS is designed to automatically seek out many "good models near good models."

In this article we highlight the relationships between MCMC methods and SSS, and describe and exemplify the differences in concept and practicalities. We provide examples in which the ability of SSS to catalog high-probability models rapidly is superior to that of competing MCMC methods, and stress the use of distributed computing to allow scaling to problems with very large numbers of predictors that otherwise would be simply infeasible.

SSS is introduced and developed in Section 2, and its use in linear and binary regression is discussed in Section 3. The conceptual basis is generic, and SSS will apply to other forms of regression models, including designed experiments with categorical covariates and interactions, but this article presents the ideas, methodology, and examples in main-effects models, in which defining model neighborhoods, a key ingredient of the methodology, is natural and interpretable. The relationship between SSS and MCMC approaches is explored in Section 4, and results from both simulation studies and real data examples that demonstrate the effectiveness of SSS are given. An example involving gene expression analysis is given in Section 5, and some concluding comments are provided in Section 6.

Chris Hans is Assistant Professor, Department of Statistics, The Ohio State University, Columbus, OH 43210 (E-mail: *hans@stat.ohio-state.edu*). Adrian Dobra is Assistant Professor, Center for Statistics and the Social Sciences and Departments of Statistics and Biobehavioral Nursing and Health Systems, University of Washington, Seattle, WA 98195 (E-mail: *adobra@stat.washington.edu*). Mike West is the Arts & Sciences Professor of Statistics and Decision Sciences, ISDS, Duke University, Durham, NC 27708 (E-mail: *mike@stat.duke.edu*). Support for this work was provided by the National Science Foundation (grants DMS-01-02227 and DMS-03-42172), the National Institutes of Health (grants NHLBI 1P01-HL-73042-02), and the W. M. Keck Foundation. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the NSF, NIH, or Keck Foundation. This work also was supported by a grant of computing time from the Ohio Supercomputer Center. The authors are grateful to Quanli Wang of Duke University for advice and assistance on computational matters and to the joint editor, associate editor, and three anonymous referees for their positive and constructive comments on the original version of the manuscript.

Journal of the American Statistical Association
June 2007, Vol. 102, No. 478, Theory and Methods
DOI 10.1198/016214507000000121

2. SEARCHING FOR REGRESSION MODELS

2.1 Shotgun Stochastic Search

The example dataset described in Section 5 has $p = 4{,}514$ possible predictor variables. If we consider regression models with up to five predictors, then there are $>10^{16}$ models; even in this constrained model space, enumeration is impossible. Denote the model space by Γ. Stochastic model search aims to discover and evaluate a (large) set of models, Γ^*, to be used in understanding model (variable subset selection) uncertainty and for prediction. Regression model shotgun stochastic search (RMSSS) is such a method. It is a regression model specific implementation of a general class of SSS methods. SSS is an iterative, local-move, neighborhood-based procedure involving three steps:

Step 1. Use the "current" model to define a neighborhood of proposal models.
Step 2. Evaluate each proposal model in this neighborhood in parallel.
Step 3. Choose a new current model from the proposals.

A key idea is that for any "current" model, there may be many other models with similar "fit" to the data—models with overlapping or collinear predictors. Quickly identifying and evaluating these models provides a rich description of part of the model space and a new set of competitive models from which to choose the next move. This generates multiple candidate models and "shoots out" proposed moves in various directions in model space.

The neighborhood of the current model must be sufficiently comprehensive to allow the search to move easily throughout the model space. This is accomplished by considering each possible predictor variable in one of the proposal models at each iteration. This approach has the added benefit that over the course of the search, every candidate variable is evaluated in the context of many different regression models. Critically, step 2 can be parallelized; each of the proposal models can be evaluated independently on separate processors, providing a clear advantage of SSS procedures over MCMC algorithms in which models are proposed and evaluated one at a time, sequentially. The criterion used to compare models is problem-specific. In a Bayesian analysis, as described here, the model "score" is posterior probability; however, the search method can be applied with other notions of model fit/score as well.

2.2 Regression Model Shotgun Stochastic Search

The two major components of SSS are the choice of neighborhood and the model move (sampling) strategy. The neighborhood component should allow consideration of each possible predictor variable at every step and, to allow the search to move freely across model size, should admit regression models of various dimensions. We take the neighborhood to be every regression model that is a one-variable change to the current model.

Let p be the total number of possible predictor variables, and let $\boldsymbol{\gamma}$ be a $p \times 1$ indicator vector with $\gamma_j = 1(0)$ if variable j is in the regression model (or not). For the moment, we consider main-effects-only regression models with continuous covariates. For a current regression model of dimension k (i.e., having k predictor variables), the neighborhood has three elements: $\text{nbd}(\boldsymbol{\gamma}) = \{\boldsymbol{\gamma}^+, \boldsymbol{\gamma}^\circ, \boldsymbol{\gamma}^-\}$, where $\boldsymbol{\gamma}^+$ is a set containing neighboring models of dimension $k+1$, called the "addition" moves; $\boldsymbol{\gamma}^\circ$ is a set containing neighboring models of dimension k, called the "replacement" moves; and $\boldsymbol{\gamma}^-$ is a set containing neighboring models of dimension $k-1$, called the "deletion" moves. Set $\boldsymbol{\gamma}^+$ contains all of the models obtained by adding any one of the $p-k$ remaining predictor variables, set $\boldsymbol{\gamma}^-$ is the k models obtained by deleting any one current variable, and $\boldsymbol{\gamma}^\circ$ is the set obtained by replacing any one current variable with any one of the $p-k$ remaining.

For example, with $p = 5$, if the current regression model is $\{x_1, x_3, x_4\}$, then

$$\boldsymbol{\gamma}^- = \big\{\{x_3, x_4\}, \{x_1, x_4\}, \{x_1, x_3\}\big\},$$

$$\boldsymbol{\gamma}^+ = \bigcup_{j\in\{2,5\}} \{x_1, x_3, x_4, x_j\}, \qquad \text{and}$$

$$\boldsymbol{\gamma}^\circ = \bigcup_{j\in\{2,5\}} \big\{\{x_1, x_3, x_j\}, \{x_1, x_j, x_4\}, \{x_j, x_3, x_4\}\big\}.$$

Note that when $2 \le k < p$, $|\boldsymbol{\gamma}^+| = p-k$, $|\boldsymbol{\gamma}^\circ| = k(p-k)$, and $|\boldsymbol{\gamma}^-| = k$, with the convention that $\boldsymbol{\gamma}^+ = \emptyset$ if $k = p$. We evaluate the null model and all possible one-variable models before starting the search, and so allow SSS to consider models only of at least dimension $k = 2$.

When p is large, $|\boldsymbol{\gamma}^\circ| \gg |\boldsymbol{\gamma}^+| \gg |\boldsymbol{\gamma}^-|$, which is problematic for sampling. If all of the models were to have equal weight and one model was sampled directly from $\text{nbd}(\boldsymbol{\gamma})$, then as $p \to \infty$, the probability of staying in the same dimension goes to $k/(k+1)$, the probability of increasing goes to $1/(k+1)$, and the probability of decreasing dimension goes to zero. To move across dimension effectively, we break sampling into two steps: three models, $\boldsymbol{\gamma}_*^+$, $\boldsymbol{\gamma}_*^\circ$, and $\boldsymbol{\gamma}_*^-$, are sampled from $\boldsymbol{\gamma}^+$, $\boldsymbol{\gamma}^\circ$, and $\boldsymbol{\gamma}^-$, and then one of the three is selected.

The (unnormalized) posterior probability, $p(\boldsymbol{\gamma}|y) \propto p(y|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$, is evaluated for each model generated in SSS. The Bayesian information criterion (BIC) can be viewed as an approximation to the marginal likelihood of a given model, $p(y|\boldsymbol{\gamma})$, under a reference prior distribution (Raftery 1995) and so could be used in similar fashion. Other scores, such as R^2 and the Akaike information criterion (AIC), can be used, but the user would have to decide how to normalize the scores into a probability distribution from which to sample. In general, we refer to a score for a model $\boldsymbol{\gamma}$ that can be normalized within a set of scores to become a probability as $S(\boldsymbol{\gamma})$.

Regression Model Shotgun Stochastic Search Schema. Let $\boldsymbol{\gamma}$ denote a regression model, and let $S(\boldsymbol{\gamma})$ be its corresponding (unnormalized) score. Given a starting model $\boldsymbol{\gamma}^{[0]}$, set $\Gamma^* = \{\boldsymbol{\gamma}^{[0]}\}$, and choose a constant B that will be the maximum number of elements of Γ^*. Iterate in $t = 1, \ldots, T$ the following steps:

Step 1. In parallel, compute $S(\boldsymbol{\gamma})$ for all $\boldsymbol{\gamma} \in \text{nbd}(\boldsymbol{\gamma}^{[t]})$, constructing $\boldsymbol{\gamma}^+$, $\boldsymbol{\gamma}^\circ$, and $\boldsymbol{\gamma}^-$. Update Γ^* to be $\Gamma^* \cup \text{nbd}(\boldsymbol{\gamma}^{[t]})$; if $|\Gamma^*| > B$, then remove the $|\Gamma^*| - B$ lowest scoring models.
Step 2. Sample $\boldsymbol{\gamma}_*^+$, $\boldsymbol{\gamma}_*^\circ$, and $\boldsymbol{\gamma}_*^-$, from $\boldsymbol{\gamma}^+$, $\boldsymbol{\gamma}^\circ$, and $\boldsymbol{\gamma}^-$, with probabilities proportional to $S(\boldsymbol{\gamma})$, normalized within each set.

Step 3. Sample $\boldsymbol{\gamma}^{[t+1]}$ from $\{\boldsymbol{\gamma}_*^+, \boldsymbol{\gamma}_*^\circ, \boldsymbol{\gamma}_*^-\}$, with probability proportional to $S(\boldsymbol{\gamma})$, normalized within this set.

Note that after running SSS, Γ^* contains the B best models evaluated, not just the best of those composing the sequence $\boldsymbol{\gamma}^{[0]}, \ldots, \boldsymbol{\gamma}^{[T]}$. That is, Γ^* contains the B best models in the collection of neighborhoods $\bigcup_{t=0}^{T-1} \text{nbd}(\boldsymbol{\gamma}^{[t]})$.

2.3 Regression Model Shotgun Stochastic Search With Higher-Order Terms

As introduced in Section 2.2 RMSSS is restricted to "main-effects-only" models. With simple modifications to the model space Γ and the neighborhood structure $\text{nbd}(\boldsymbol{\gamma})$, RMSSS can accommodate more complicated regression settings, such as the inclusion of higher-order terms. Consider first the use of power terms where for each predictor variable x_{ij} (the jth covariate for the ith individual), the terms x_{ij}^d, $d = 1, \ldots, D$, are of interest. In a case in which the investigator will allow any of these terms to enter freely in a model, the model space can be redefined so that $\boldsymbol{\gamma}$ is now a $Dp \times 1$ vector, and the $(j + (d-1)p)$th element of $\boldsymbol{\gamma}$, $j = 1, \ldots, p$, corresponds to x_j^d. RMSSS then proceeds as described in Section 2.2.

Now consider the situation in which the investigator wishes to allow for the inclusion of interaction (including squared) terms, $x_{ij}x_{ik}$, $1 \le j \le k \le p$, but also wishes to impose the constraint that if $x_{ij}x_{ik}$ is in a model, then both main effects (or the single main effect in the case of a squared term) also must be in the model. To perform an SSS over this model space, we redefine the neighborhood structure for a given model with specified main-effects and interaction terms as follows. The deletion set, $\boldsymbol{\gamma}^-$, is obtained by first deleting, one at a time, the second-order terms, and then deleting, one at a time, the main-effects terms, with the proviso that when a main effect is deleted, all of the second-order terms in which it is involved are also deleted. The addition set, $\boldsymbol{\gamma}^+$, is obtained by first adding, one at a time, each of the (main-effects) variables not currently in the model. Then, for each of these $p - k$ models, the "added" variable is interacted, one at a time, with each of the main-effects variables currently in the model, including the added variable itself for the squared term. The replacement set, $\boldsymbol{\gamma}^\circ$, is obtained in a similar manner by swapping each current main effect with each of the main effects not currently in the model, deleting higher-order terms when necessary, and interacting each "swapped in" variable with each of the remaining main effects. For example, if $p = 3$ and the current model is $\{x_1, x_2, x_1^2\}$, then the neighborhood would be

$$\boldsymbol{\gamma}^\circ = \left\{\{x_1, x_3, x_1^2\}, \left\{\bigcup_{j\in\{1,3\}} \{x_1, x_3, x_1^2, x_jx_3\}\right\}, \{x_2, x_3\}, \left\{\bigcup_{j\in\{2,3\}} \{x_2, x_3, x_jx_3\}\right\}\right\},$$

$$\boldsymbol{\gamma}^- = \left\{\{x_1, x_2\}, \{x_2\}, \{x_1, x_1^2\}\right\},$$

$$\boldsymbol{\gamma}^+ = \left\{\{x_1, x_2, x_3, x_1^2\}, \bigcup_{j\in\{1,2,3\}} \{x_1, x_2, x_3, x_1^2, x_jx_3\}\right\}.$$

3. LINEAR AND BINARY REGRESSION

3.1 Normal Linear Regression

Consider the normal linear regression model $\mathbf{Y} \sim \text{N}(X\boldsymbol{\beta}, \sigma^2 I_n)$, where $\mathbf{Y}$ is an $n \times 1$ response variable, $X = (x_1, \ldots, x_n)'$ is an $n \times p$ data matrix for the n samples, and the x_i's are $p \times 1$ vectors of covariate information. Assume that both the x and y data are centered and normalized, so that we do not include an intercept term in the model. We assume priors on $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)'$ that are consistent across models in the sense that they are derived from an encompassing model through conditioning. Following Dobra et al. (2004) and Geiger and Heckerman (2002), we take $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2, \boldsymbol{\gamma}) = \text{N}(0, \tau^{-1}\sigma^2 I_k)$ and $p(\sigma^2|\boldsymbol{\gamma}) = \text{IG}((\delta + k)/2, \tau/2)$, where $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the vector of regression coefficients under a model $\boldsymbol{\gamma}$ with k variables. Because we have scaled the data to have unit variance, we typically set $\tau = 1$ to reflect this common scale. The prior distribution for σ^2 has a finite first moment for all $k \ge 0$ when $\delta > 2$, so we typically set $\delta = 3$. To find $p(\boldsymbol{\gamma}|y)$, we first compute the marginal likelihood $p(y|\boldsymbol{\gamma}) = \int p(y|\boldsymbol{\theta}, \boldsymbol{\gamma})p(\boldsymbol{\theta}|\boldsymbol{\gamma})\, d\boldsymbol{\theta}$, which has a closed-form solution under the foregoing formulation (see Dobra et al. 2004). Then, by Bayes' theorem, the posterior probability of any model is $p(\boldsymbol{\gamma}|y) \propto p(y|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$.

3.2 Binary Regression

In the case of independent binary outcomes, y_i, consider the logistic regression $p(y|\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^n p_i^{y_i}(1 - p_i)^{1-y_i}$, where $p_i = 1/(1 + \exp\{-(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}_{\boldsymbol{\gamma}})\})$ and $\mathbf{x}_i$ contains only those variables indicated by the model $\boldsymbol{\gamma}$. Note the inclusion of the intercept term β_0, which is necessary to account for the baseline response probability. We set $p(\beta_0, \boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) = \text{N}(0, \tau I_{k+1})$, where $k = \sum_{j=1}^p \gamma_j$, and assuming standardized predictor variables, we typically take $\tau = 1$ to place appropriate prior mass on reasonable values of the regression coefficients.

The marginal likelihood is not available in closed form but can be approximated through the Laplace approximation $\hat{p}(y|\boldsymbol{\gamma}) = (2\pi)^{p/2}|\hat{\Sigma}|^{1/2}p(y|\hat{\boldsymbol{\beta}}, \boldsymbol{\gamma})p(\hat{\boldsymbol{\beta}}|\boldsymbol{\gamma})$ (DiCiccio, Kass, and Wasserman 1997), where we find $\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} p(y|\boldsymbol{\beta}, \boldsymbol{\gamma})p(\boldsymbol{\beta}|\boldsymbol{\gamma})$ through Newton's method and compute $\hat{\Sigma}$ as the inverse of the negative Hessian matrix of $\log[p(y|\boldsymbol{\beta}, \boldsymbol{\gamma})p(\boldsymbol{\beta}|\boldsymbol{\gamma})]$ evaluated at $\hat{\boldsymbol{\beta}}$.

3.3 Prior Over the Model Space

As dimension increases, it is critical to use priors over model dimension that encourage sparsity, because large models are often less interpretable and there is a risk of overfitting when n is small relative to p. Here we use the standard model selection prior $p(\boldsymbol{\gamma}) = \pi^k(1 - \pi)^{p-k}$, where k is the number of variables in the model and π is a hyperparameter representing the probability that a variable is in the model (with all variables treated exchangeably). This induces a binomial prior distribution over model size, and thus the prior expected model size is $p\pi$. We typically set $\pi = k'/p$, with k' small relative to p, to encourage sparsity.

3.4 Variable Identification and Prediction

Measure the relative importance of individual models $\gamma \in \Gamma^*$ by

$$\tilde{p}(\boldsymbol{\gamma}|y) = C^{-1}p(y|\boldsymbol{\gamma})p(\boldsymbol{\gamma}), \tag{1}$$

where the norming constant is the posterior mass contained in Γ^*, $C = \sum_{\boldsymbol{\gamma}\in\Gamma^*} p(y|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$. If we could have explored the entire space (so that $\Gamma^* = \Gamma$), then (1) would represent the posterior probability of model $\boldsymbol{\gamma}$. But because we explored only some part of the model space, (1) represents the posterior probability of model $\boldsymbol{\gamma}$ *conditioned* on the set Γ^*. View $\tilde{p}(\boldsymbol{\gamma}|y)$ as a measure of relative importance of model $\boldsymbol{\gamma}$ in the context of the top predictive models. Similarly, measure the relative importance of variable x_j by

$$\tilde{p}(\gamma_j = 1|y) = \sum_{\boldsymbol{\gamma}\in\Gamma^*} \mathbf{1}_{\{\gamma_j=1\}}\tilde{p}(\boldsymbol{\gamma}|y). \tag{2}$$

Predictions can now be based on the set of top models, representing a conditional posterior.

4. THE NATURE AND EFFECTIVENESS OF SHOTGUN STOCHASTIC SEARCH

4.1 Relationship to Markov Chain Monte Carlo

In cases of high-dimensional parameter spaces, MCMC approaches are often used not with the aim of performing Monte Carlo integration to summarize the posterior distribution, but rather as a stochastic search tool to identify regions of high posterior probability (or, in the context of model selection, to identify the "best" models). In this section we show that small changes to SSS result in an MCMC algorithm of a fundamentally different from than common MCMC approaches.

Consider use of a Metropolis–Hastings algorithm to sample from a discrete distribution, $P(x)$, where we can evaluate $P(x)$ up to a normalizing constant, $P(x) \propto Q(x)$. Consider proposal distributions that sample from $P(x)$ restricted to a neighborhood $N(\cdot)$,

$$T(x'; x_t) = \frac{P(x')\mathbf{1}(x' \in N(x_t))}{\sum_{s\in N(x_t)} P(s)} = \frac{Q(x')\mathbf{1}(x' \in N(x_t))}{\sum_{s\in N(x_t)} Q(s)}.$$

As long as we start the chain in a region of nonzero probability, the acceptance probability at each iteration is

$$\alpha = \min\left\{1, \frac{\sum_{s\in N(x_t)} Q(s)}{\sum_{s\in N(x')} Q(s)}\right\}. \tag{3}$$

We can easily adapt the SSS algorithm described in Section 2.2 to become a Metropolis–Hastings algorithm using the proposal distribution described earlier. Relating notation, we have that $P(x_t)$ is $p(\boldsymbol{\gamma}^{[t]}|y)$, $Q(x_t)$ is $S(\boldsymbol{\gamma}^{[t]}) = p(y|\boldsymbol{\gamma}^{[t]})p(\boldsymbol{\gamma}^{[t]})$, and $N(x_t)$ is $\text{nbd}(\boldsymbol{\gamma}^{[t]})$. After performing step 1 at iteration t in SSS, sample a proposal $\boldsymbol{\gamma}'$ from the discrete distribution $S(\cdot)$ normalized within $\text{nbd}(\boldsymbol{\gamma}^{[t]})$, and set $\boldsymbol{\gamma}^{[t+1]} = \boldsymbol{\gamma}'$ with probability α from (3); otherwise, set $\boldsymbol{\gamma}^{[t+1]} = \boldsymbol{\gamma}^{[t]}$. Steps 2 and 3, which are related to the two-stage sampling process that corrects the dimensional imbalance, are ignored.

The form of the acceptance probability (3) indicates that Metropolized SSS behaves much differently than standard MCMC approaches such as the MCMC model composition (MC^3) algorithm of Madigan and York (1995) and Raftery et al. (1997), and the related approach of Brown et al. (1998a). MC^3 constructs a Markov chain over the model space by first defining a neighborhood $\text{nbd}_*(\boldsymbol{\gamma}^{[t]}) = \boldsymbol{\gamma}^+ \cup \boldsymbol{\gamma}^- \cup \boldsymbol{\gamma}^{[t]}$, using our notation from Section 2.2. A proposal distribution T_* is then defined by setting $T_*(\boldsymbol{\gamma}'; \boldsymbol{\gamma}^{[t]}) = 0$ for all $\boldsymbol{\gamma}' \notin \text{nbd}_*(\boldsymbol{\gamma}^{[t]})$ and setting $T_*(\boldsymbol{\gamma}'; \boldsymbol{\gamma}^{[t]})$ constant for all $\boldsymbol{\gamma}' \in \text{nbd}(\boldsymbol{\gamma}^{[t]})$. As the MC^3 algorithm proceeds, if the chain is in state $\boldsymbol{\gamma}^{[t]}$, then a proposed move $\boldsymbol{\gamma}'$ is drawn from $T_*(\boldsymbol{\gamma}'; \boldsymbol{\gamma}^{[t]})$, a discrete uniform distribution over $\text{nbd}(\boldsymbol{\gamma}^{[t]})$. The proposed move is accepted with probability

$$\alpha_* = \min\left\{1, \frac{p(y|\boldsymbol{\gamma}')p(\boldsymbol{\gamma}')}{p(y|\boldsymbol{\gamma}^{[t]})p(\boldsymbol{\gamma}^{[t]})}\right\}.$$

Effectively, at each iteration, MC^3 randomly chooses a component of $\boldsymbol{\gamma}^{[t]}$ and switches its value of 0/1 with probability α_*.

The acceptance probability α_* depends only on $\boldsymbol{\gamma}^{[t]}$ and $\boldsymbol{\gamma}'$ and favors rejecting moves to lower-probability models. However, the acceptance probability α for Metropolized SSS, depends on the amount of posterior mass in the *neighborhoods* around $\boldsymbol{\gamma}^{[t]}$ and $\boldsymbol{\gamma}'$, and favors moves to models away from local modes, discouraging entrapment in particular regions of model space.

4.2 Comparison With Markov Chain Monte Carlo Methods

Two popular MCMC approaches for model space exploration are MC^3 (described earlier) and Gibbs sampling. George and McCulloch (1997) and Smith and Kohn (1996, 1997) described how to construct Gibbs samplers over a model space in the conjugate setting where $p(y|\boldsymbol{\gamma})$ is available in closed form. A one-at-a-time, fixed-scan Gibbs sampler creates a sequence of models $\boldsymbol{\gamma}^{[1]}, \boldsymbol{\gamma}^{[2]}, \ldots$, by updating the components of $\boldsymbol{\gamma}$ by sampling from $p(\gamma_j|\boldsymbol{\gamma}_{-j}, y) \propto p(y|\boldsymbol{\gamma})p(\gamma_j|\boldsymbol{\gamma}_{-j})$ for $j = 1, \ldots, p$ at each iteration.

We implemented SSS and both MC^3 and the Gibbs sampler for the Keck dataset described in Section 4.4, using the observed rather than the simulated data. In both cases, we used the sparsity-inducing prior $\pi = 10/p$. SSS was run for 40,000 iterations, which resulted in a total of 1,137,195,208 total model evaluations; we then ran the Gibbs sampler, which evaluates $p = 8{,}408$ models per iteration, for 135,252 iterations, and ran MC^3, which evaluates one model per iteration, for 1,137,195,208 iterations, so that the three runs represent approximately the same number of model evaluations. Figure 1 shows the accumulated posterior mass in the set of the top 1 million models for each run as a function of the number of models evaluated, with the plot normalized by the total mass found by SSS. By the end of the run, the Gibbs sampler and MC^3 had accumulated 97.49% and 92.95% of the total mass accumulated by SSS. From the plot, however, we see that SSS dominated in its ability to accumulate posterior mass with fewer model evaluations.

We note also that SSS was used in analyses of this dataset by Rich et al. (2005) to define gene expression–based regression predictions of survival among brain cancer patients. Multiple models involving three to five predictors were identified in that analysis, and the leave-one-out cross-validation predictions based on model averaging across many "top models" suggests

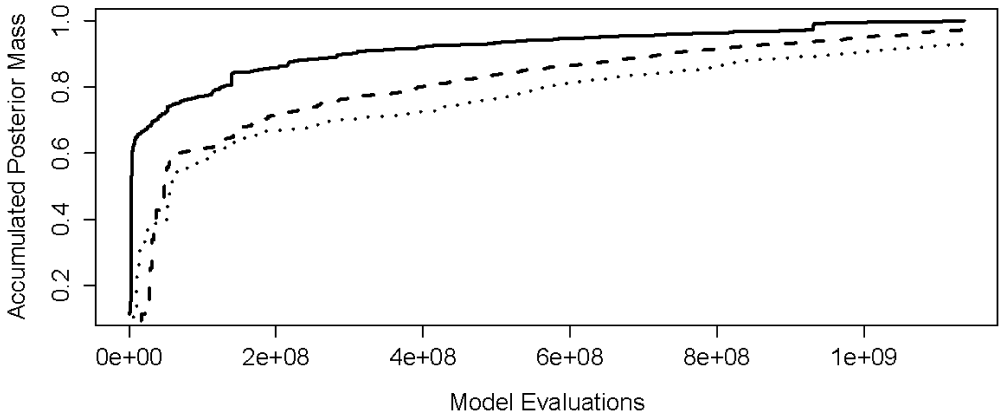

Figure 1. Accumulated Posterior Mass for SSS (——), Gibbs (- - - - -), and MC^3 (·····).

a reliable predictive relationship. Moreover, the most probable models generally involve at least two of three biologically interpretable genes, with other top models involving genes with expression profiles highly correlated with these. The "top models" identified by SSS in that application were not discovered using several variants of stepwise search.

4.3 The Null Model Scenario

It is of interest to examine how SSS performs in the situation in which the generating model is the null normal linear model. We simulated a response vector of length $n = 100$ from independent standard normal distributions, and for each observation independently simulated $p = 5{,}000$ covariates from independent standard normal distributions. The data were mean-centered and standardized as in Section 3.1, and the best 1 million models from a run of SSS were recorded with $k' = 10$ in the model space prior. Model-averaged fitted values are shown in Figure 2. If the data were fit perfectly by the (weighted average of the) models, then the points would fall on the line $y = x$; rather, we see considerable shrinkage toward a mean of 0, indicating dominance of the null model in Γ^*. Indeed, the null model was the highest scoring model found by SSS.

4.4 Simulated Data Example

In this section we report a simulation study based on a real dataset to demonstrate the effectiveness of SSS as the number of possible predictor variables increases. Here we do not restrict ourselves to a fixed-dimensional SSS as in the previous two sections. The data on which we base the simulation is a gene expression dataset from a survival study in brain cancer based at the W. M. Keck Center for Neuro-Oncology at Duke University. A detailed description of the data, along with an initial analysis, was given by Rich et al. (2005).

The study group consists of 41 patients, each of which has gene expression data consisting of 8,408 genes from a tumor specimen. We selected four genes from the dataset as the variables composing the "true" model $\boldsymbol{\gamma}^*$ and simulated $m = 1, \ldots, 50$ outcomes using the actual gene expression values x_{ij} for the $j = 1, \ldots, 4$ "true" variables according to the regression model $y_i^{(m)} = 1.3x_{i1} + .3x_{i2} - 1.2x_{i3} - .5x_{i4} + \varepsilon_i^{(m)}$, for $i = 1, \ldots, 41$, where the $\varepsilon_i^{(m)}$'s are iid mean-0 normal random variables with variance .5. The simulated outcomes were then standardized to have mean 0 and unit variance within each of the 50 simulations. Some information is known about the four genes that we chose to compose the true model: one is from the RAS oncogene family, one is a glioblastoma-amplified

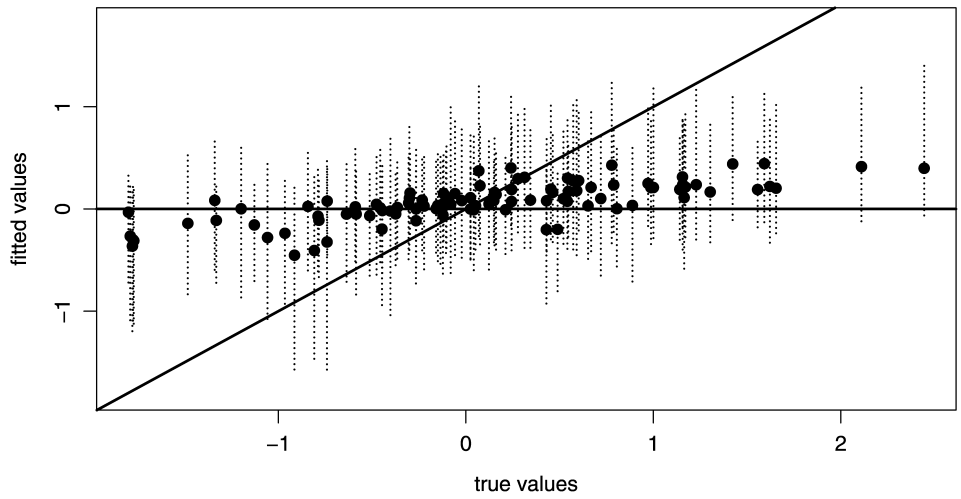

Figure 2. Model-Averaged Fitted Values When the True Model Is the Null Model, With 95% Intervals.

sequence, one is a leukocyte protease inhibitor, and one is a CAT56 protein. These genes were chosen in part because they exhibit some correlation with other genes in the dataset. The regression coefficients were chosen so that two were large and two were small relative to the variance of ε_i.

To assess the performance of SSS with increasing dataset size, we ran SSS for the 50 simulated responses using datasets with increasing values of p, as shown in Figure 3. The datasets were constructed by first reordering the observed $41 \times 8{,}408$ data matrix X so that the four variables used in the simulation were designated variables 1, 2, 3, and 4. To construct a data matrix $X^{(m,p)}$ for a particular simulation, when $p \leq 8{,}408$, we extracted the first p columns of X to form $X^{(m,p)}$ and then randomly permuted the columns. Hence all 50 datasets $X^{(m,p)}$ for a given $p < 8{,}408$ contain the same variables and differ only by a column permutation. For the datasets with $p > 8{,}408$, before permuting the columns, we added $p - 8{,}408$ columns of random draws from a $N(0, I_{41})$ distribution (after centering and scaling the random draws), effectively adding random noise to the dataset.

Prior distributions over the parameter space in the simulation study are consistent with those used in the analysis by Rich et al. (2005), with $\tau = 1$ and $\delta = 3$ as described in Section 3.1. For the model space prior, we set $\pi = 4/p$ as in Section 3.3 to maintain focus on sparse models as p increases. We note that the general conclusions and features of this (and other) examples remained similar as we reran the analysis using different values of π, although it is of course critical that π be small, to ensure a focus on relatively small models. Inference on π itself in the SSS context is an open question.

For a given run of SSS, we declared that SSS had found the true model when the true model was evaluated [i.e., when $\boldsymbol{\gamma}^* \in \text{nbd}(\boldsymbol{\gamma}^{[t]})$]. For each value (m, p), if SSS found the true model within 10,000 iterations, then we recorded the number of iterations required to find the model and the elapsed time. If the model was not found within 10,000 iterations, then we recorded the time required for the 10,000 iterations.

Computation was done using 21 processing elements (1 head node and 20 compute nodes) on a cluster of dual-processing, 3.1-GHz Intel x86-based machines running Linux. SSS was run for one value of (m, p) at a time using the 21 processors. Results from the simulation study are shown in Figure 3. Although increasing the number of irrelevant variables in the dataset resulted in an increased number of iterations needed to find the true model, the true model was still found by SSS a large percentage of the time.

4.5 Fixed-Dimensional Shotgun Stochastic Search for Orthogonal Designs

Another metric by which we can compare SSS with an MCMC method is to examine the expected number of iterations until the "true" model is found. Consider a fixed-dimensional SSS, where we condition on a particular number of variables k and allow moves only within this dimension, effectively setting $\text{nbd}(\boldsymbol{\gamma}) = \boldsymbol{\gamma}^\circ$. A fixed-dimensional SSS creates a Markov chain $\{\boldsymbol{\gamma}^{[t]}\}$ over the state space of models restricted to size k. Assuming that the true model $\boldsymbol{\gamma}^*$ is of dimension k, we can define a mapping $Z_t \equiv \psi(\boldsymbol{\gamma}^{[t]}) = k - \sum_{j=1}^{p} \boldsymbol{\gamma}_j^* |\boldsymbol{\gamma}_j^* - \boldsymbol{\gamma}_j^{[t]}|$, where $Z_t \in \{0, \ldots, k\}$ indicates how many of the variables in $\boldsymbol{\gamma}_t$ are shared by $\boldsymbol{\gamma}^*$. Under certain restrictions on the model space (described later), the one-step transition probabilities for the induced Markov chain $\{Z_t\}$ can be easily computed, and the expected number of iterations until the true model $\boldsymbol{\gamma}^*$ is found can be evaluated.

Consider the case in which all possible predictor variables are orthogonal, that is, where $x_i' x_j = 0$ for all $i \neq j$. It can be shown (see the on-line technical report) that under the model specification in Section 3.1, the marginal likelihood for a given model is simply a function of the least squares estimates of the regression coefficients. Moreover, the one-step transition probabilities for the induced Markov chain $\{Z_t\}$ can be computed by making the further assumption that all variables not in the true model have the same (relatively small) scaled regression coefficient $\epsilon = x_j' y$, and that all variables that are in the true model have the same (relatively large) scaled regression coefficient $\lambda = x_j' y$.

The solid lines in Figure 4(a) show the expected number of steps until the true model is evaluated (on the $\log_{10}$ scale) as a function of p for values $k = 3, 4, 5, 6$ under the foregoing assumptions. Here we took $n = 500$ and set $\epsilon = (n-1).005$, $\lambda = (n-1).1$, $\tau = 1$, and $\delta = 3$. The dashed lines provide a comparison to a fixed-dimensional MCMC search; each step in

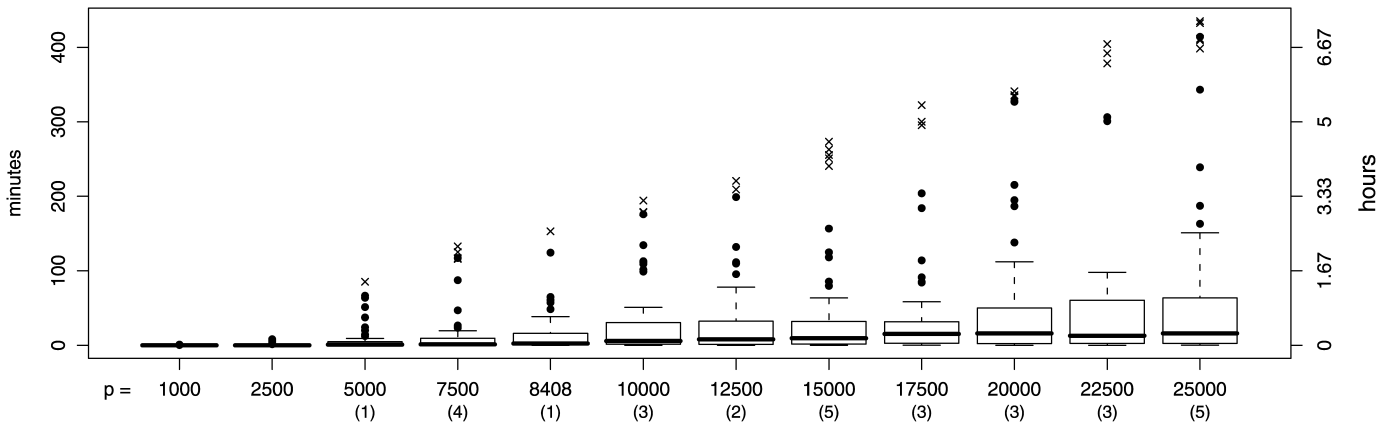

Figure 3. Time Required to Find the True Model for the Simulation Study Described in Section 4.4. The numbers in parentheses indicate the number of models not found by SSS in 10,000 iterations for a given dataset size p, denoted by × in the plot. The boxplots are based only on runs for which SSS found the true model.

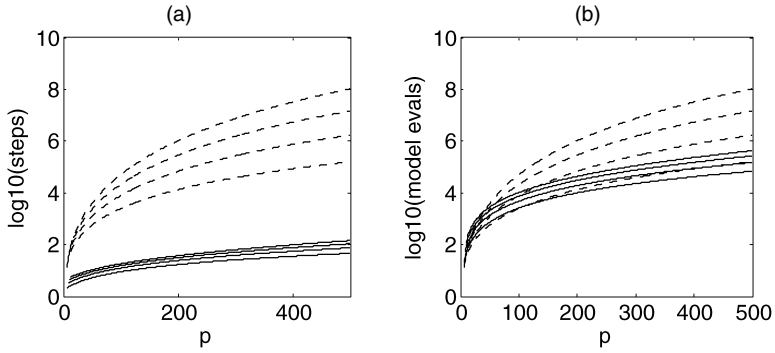

Figure 4. Comparison of Fixed-Dimensional SSS and MCMC (Metropolis) Approaches Under the Assumptions Given in Section 4.5. (a) Base 10 logarithms of the average number of steps needed before the true model is found, as a function of p; (b) Base 10 logarithms of the average number of model evaluations needed before the true model is found, as a function of p. (——, SSS; - - - - -, MCMC.)

the chain proceeds by randomly proposing to replace one variable in the current model with one variable not in the current model and then accepting the proposed move with probability computed through the Metropolis ratio. SSS clearly dominates with respect to the number of steps needed on average before the true model is evaluated; however, this comparison is unfair from a computational perspective, because fixed-dimensional SSS evaluates $k(p-k)$ models per step, whereas the Metropolis algorithm evaluates only one model per step. A more level comparison between the two methods is provided in Figure 4(b), which shows the expected number of model evaluations until the true model is evaluated. SSS is competitive with MCMC for small p and dominates as p grows larger.

5. SHOTGUN STOCHASTIC SEARCH EXAMPLE USING GENE EXPRESSION DATA

5.1 Data and Prediction Analysis Context

The dataset comprises coupled gene expression and lymph node positivity status in human breast cancers. From a database of about 350 cases, we identified those clinically defined as low risk for disease recurrence or death from disease in terms of lymph node negativity (no evidence of cancer metastasis in the axillary lymph nodes) at the point of surgery; we compared these patients with those in a generally much higher-risk group (i.e., those with at least nine nodes in the axillary regions showing evidence of cancer metastasis). This analysis follows previous work and relates to the general interest in the potential for tumor-derived gene expression profiles to aid in prognosis—in this case, improved prediction of low risk versus high risk based on genomic information could affect decisions about postsurgical treatments (West et al. 2001; Huang, West, and Nevins 2002; Nevins et al. 2003; Huang et al. 2003; Pittman et al. 2004). Prediction of lymph node status based on gene expression profiles is a challenging problem, due to the complex heterogeneity of the disease in terms of genetic/genomic and environmental factors, and also due to the levels of experimental and technical noise in gene expression data. Advances in our ability to better predict the state would be of substantial interest in clinical cancer genomics.

The data consist of $n = 148$ samples with $n_0 = 100$ low-risk (node-negative) and $n_1 = 48$ high-risk (high node-positive) cases. Gene expression data are available on Affymetrix HU95aV2 oligonucleotide microarrays, which were processed using the current standard RMA method (Irizarry et al. 2003a,b) to generate summary estimates of expression levels of each gene in each sample. These primary RMA data were then further screened and normalized, and we selected a total of 4,512 genes showing evidence of more than trivial variation above the noise levels. In addition to these candidate predictors, some traditional clinical factors are available for each patient, including estimated tumor size (in centimeters) and protein assay–based estrogen receptor (ER) status, coded as a binary covariate. Using the gene expression data together with these two clinical factors thus provides $p = 4{,}514$ candidate predictors.

5.2 Small Subsets Regression Analysis

We use binary regression models as described in Section 3.2 to assess the relative importance of the individual genes and clinical factors in the context of lymph node invasion and to serve as a predictive model. We take $y_i = 0$ to denote node negative cases and $y_i = 1$ to denote advanced nodal metastasis. The full vector of covariates, $\mathbf{x}_i$, is a $4{,}515 \times 1$ vector consisting of an intercept term, the two clinical variables (tumor size and ER status), and the 4,512 gene expression variables.

Because our focus is on sparse models, we take the prior distribution over the model space to be as described in Section 3.3 with $\pi = 10/4{,}514$. For $p(\boldsymbol{\beta}|\boldsymbol{\gamma})$, we take $\tau = 1$ as described in Section 3.2. After running SSS for 20,000 iterations and saving the top 100,000 models evaluated, we combined the results through the model-averaging techniques outlined in Section 3.

5.3 Results

The top 100,000 models evaluated contain a mix of between one and seven variable models, as shown in Table 1. We compute a measure of posterior importance of model size, $|\boldsymbol{\gamma}|$, as $\tilde{p}(|\boldsymbol{\gamma}| = k|y) = \sum_{\boldsymbol{\gamma}\in\Gamma^*} \mathbf{1}_{\{|\boldsymbol{\gamma}|=k\}}\tilde{p}(\boldsymbol{\gamma}|y)$, where $|\boldsymbol{\gamma}|$ in the context of binary regression refers to the number of predictors in the model minus the intercept term. Under our model specification, the data give the most support to small subset regressions of size five, six, and four, in that order. SSS did not find any models of size eight or greater to belong in the list of top models.

Conditionally on Γ^*, eight genes were found to have a posterior importance measure (2) $>.10$, as shown in the diagonal entries of Table 2. These genes dominate the list of models, because most of the four-, five-, and six-variable models include some subset of these genes. The most important

Table 1. Model Size (k) Importance Measures, Conditioned on the Top 100,000 Models Discovered by SSS

	k						
	1	*2*	*3*	*4*	*5*	*6*	*7*
Number of models	1	54	1,311	11,838	54,597	30,619	1,580
$\tilde{p}(\|\boldsymbol{\gamma}\| = k\|y)$	$<.001$	.001	.020	.184	.534	.253	.007

gene, RGS3, occurs in almost all of the models. We also computed pairwise importance measures according to $\tilde{p}(\gamma_i = \gamma_j = 1|y) = \sum_{\boldsymbol{\gamma}\in\Gamma^*} \mathbf{1}_{\{\gamma_i=\gamma_j=1\}}\tilde{p}(\boldsymbol{\gamma}|y)$. These values, reported for the top eight variables in the off-diagonal entries of Table 2, confirm that models consisting of the top four variables dominate the list. Indeed, the four-way inclusion probability, conditional on Γ^*, for the top four variables is .244, just less than one-third of the total mass for models with $k = 5, 6, 7$.

To assess the fit of the model, we computed model-averaged mean probabilities p_i and associated 80% intervals using the top 10 models. Figure 5 plots these model-averaged fitted values versus the linear predictor $\log(p_i/(1 - p_i))$, which serves as a linear risk index. The fitted values have been corrected for the baseline incidence rate of 32%, so that .5 provides a reference point. As is well known, logistic models adapt to the empirical base rate (here about 32%) in estimating the intercept of the regression, so the resulting predictive probabilities have the interpretation as posterior probabilities relative to an implicit prior probability of about .32. To reference any other base rate, say $\Pr(y_i = 1) = a$, we can simply adjust p_i to p_i^a, where $p_i^a/(1 - p_i^a) = a(1 - .32)p_i/((1 - a).32(1 - p_i))$. We generally prefer to present predictions referenced to $a = .5$ unless there is substantive prior information about a scientifically relevant base rate. The model fit is quite good: 96% of the positives are $>.5$, and 89% of the negatives are $<.5$.

To assess how the top genes combine across models in a predictive context, we took the genes composing the top 10 models (a total of 18 genes) and created 2 "metagenes," the first 2 principal components from a singular value decomposition of the (mean-centered) 18 genes. Figure 6 shows the association between these two metagenes and the model-averaged linear predictor computed earlier. We expect to find a concordance between this empirical metagene and the averaged predictions, but it is evident from the variation in the scatterplot that the complex, data-weighted mixing over the set of regression models generates predictions that are not simply captured by a single linear fit—the metagene—to the selected set of most interesting predictor variables.

One key interest in model averaging in the face of regression model uncertainty is the consequent robustness generally realized in out-of-sample prediction as a result. Selecting one model, as in typical applications of stepwise procedures, will almost surely underestimate predictive uncertainties and lead to less robust and reliable out-of-sample predictions. Prediction is also the key tool in model assessment, and we use it here to assess aspects of the current model fit. To do this, we perform a leave-one-out cross-validation prediction analysis. Leaving out observation i, we recompute model-probabilities and derive model-averaged predictions of the response probability for case i based on the remaining observations. A histogram of predicted risk index, $\log(p_i/(1 - p_i))$, is shown in Figure 7. On the basis of simple thresholding of these point estimates at 0, corresponding to a simple thresholding of the corresponding point predictions of metastasis, the analysis indicates an approximate sensitivity of 79.2% and a specificity of 76%. This level of predictive discrimination is quite high and suggests promise for the approach relative to previous analyses on much smaller and selected subsets of patients (West et al. 2001; Huang et al. 2003).

6. ADDITIONAL COMMENTS

We have presented a novel stochastic search approach for exploring regression model spaces using the power of distributed computing to allow consideration of potentially tens of thousands of possible predictor variables. The SSS approach is quite general and, in addition to the linear and binary regression models that we have considered here, can be applied to any generalized linear regression model as long as the marginal likelihood can be evaluated or approximated. SSS methods also can handle generalized linear modeling frameworks in which there is uncertainty in the choice of link function; consider SSS for binary regression where $\text{nbd}(\boldsymbol{\gamma})$ is doubled in size by computing each model under both a probit link and a logit link. Some current and recent analyses demonstrate the ability of SSS to rapidly identify multiple regions of model space exhibiting high posterior probability—or, more generally, high model "scores"—and the utility of the approach in contexts in

Table 2. Genewise and Pairwise Importance Measures

	RGS3	*DXYS155E*	*ATP6V1F*	*MGC8721*	*VDAC1*	*GEM*	*WSB1*	*PRRG1*
RGS3	.991	.716	.495	.351	.169	.133	.125	.110
DXYS155E		.716	.454	.319	.069	.069	.121	.101
ATP6V1F			.498	.250	.010	.045	.108	.039
MGC8721				.352	.016	.042	.054	.006
VDAC1					.171	.037	.001	.047
GEM						.134	*	*
WSB1							.125	.002
PRRG1								.110

NOTE: The diagonal elements are the quantities $\tilde{p}(\gamma_i = 1|y)$, and the off-diagonal elements are the quantities $\tilde{p}(\gamma_i = \gamma_j = 1|y)$. The character "*" indicates a value $<.001$.

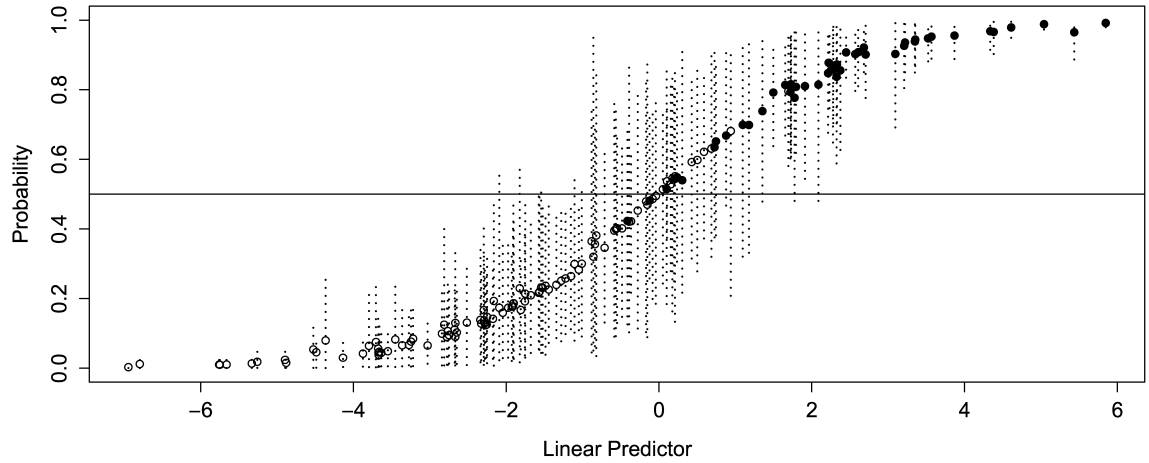

Figure 5. Model-Averaged Fitted Values Based on the Top 10 Models, With Associated 80% Intervals. The solid circles represent $y_i = 1$; the open circles, $y_i = 0$.

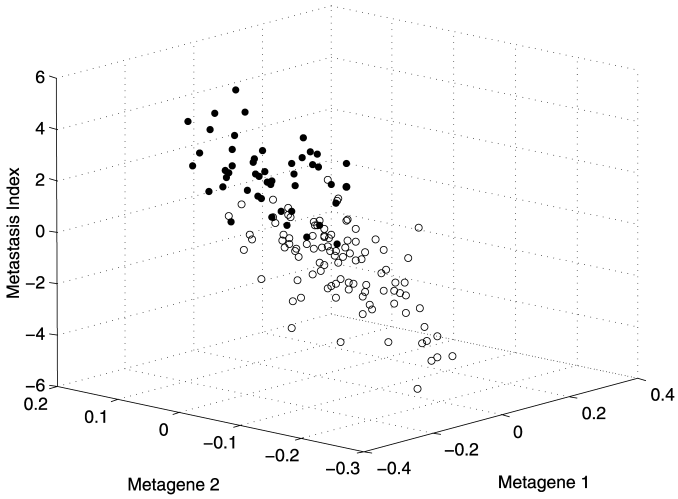

Figure 6. Association of Metagenes With the Model-Averaged Metastasis Index (linear predictor) Based on the 18 Genes Composing the Top 10 Models. Solid circles represent $y_i = 1$; open circles, $y_i = 0$.

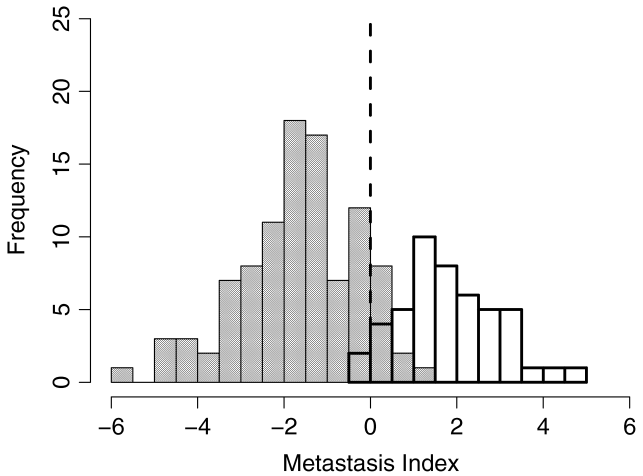

Figure 7. Histograms of the Leave-One-Out Cross-Validated Predictions on the Linear Predictor Scale. The shaded histogram represents the true negatives; the heavy-outlined histogram, the true positives.

which traditional search and MCMC methods are simply ineffective due to both dimension and the subtlety of predictive relationships in the context of noise and complex patterns of collinearity. Two recent examples in cancer genomics studies, one using linear regression (Rich et al. 2005) and one using logistic regression (Dressman et al. 2006), have illustrated this in connection with both predictive utility and variable selection/identification in challenging contexts.

We note that applications outside of regression are possible as well. For example, Jones et al. (2005) considered SSS in the context of Gaussian graphical model determination, and we anticipate further developments in that area as well as others. Some topics of current interest include improved computational implementation and more general classes of prior distributions over model spaces.

Software implementing the SSS analysis for linear regression, binary regression using logistic models, and survival regression using Weibull models is available for use by interested readers. The code is written in C++ and uses MPI for implementation in a distributed Beowulf cluster environment. Executable files for a serial implementation of SSS are also available, and the code may be modified to implement other sampling distributions. Full details are available at *www.stat.osu.edu/~hans/sss/* or at *www.isds.duke.edu/research/software* under the SSS item listing.

[Received June 2005. Revised October 2006.]

REFERENCES

Brown, P. J., Vannucci, M., and Fearn, T. (1998a), "Bayesian Wavelength Selection in Multicomponent Analysis," *Journal of Chemometrics*, 12, 173–182.

——— (1998b), "Multivariate Bayesian Variable Selection and Prediction," *Journal of the Royal Statistical Society*, Ser. B, 60, 627–641.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998), "Bayesian CART Model Search," *Journal of the American Statistical Association*, 93, 935–948.

DiCiccio, T. J., Kass, R. E., and Wasserman, L. (1997), "Computing Bayes Factors by Combining Simulation and Asymptotic Approximations," *Journal of the American Statistical Association*, 92, 903–915.

Dobra, A., Hans, C., Jones, B., Nevins, J., Yao, G., and West, M. (2004), "Sparse Graphical Models for Exploring Gene Expression Data," *Journal of Multivariate Analysis*, 90, 196–212.

Dressman, H. K., Hans, C., Bild, A., Olson, J. A., Rosen, E., Marcom, P. K., Liotcheva, V., Jones, E., Vujaskovic, Z., Marks, J., Dewhirst, M. W., West, M., Nevins, J. R., and Blackwell, K. (2006), "Gene Expression Profiles of Multiple Breast Cancer Phenotypes and Response to Neoadjuvant Chemotherapy," *Clinical Cancer Research*, 12, 819–826.

Furnival, G. M., and Wilson, R. W. (1974), "Regression by Leaps and Bounds," *Technometrics*, 16, 499–511.

Geiger, D., and Heckerman, D. (2002), "Parameter Priors for Directed Acyclic Graphical Models and the Characterization of Several Probability Distributions," *The Annals of Statistics*, 30, 1412–1440.

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

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

——— (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Geweke, J. (1996), "Variable Selection and Model Comparison in Regression," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 609–620.

Green, P. J. (1995), "Reversible-Jump Markov Chain Monte Carlo-Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Hocking, R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1–49.

Huang, E., Chen, S., Dressman, H., Pittman, J., Tsou, M. H., Horng, C. F., Bild, A., Iversen, E. S., Liao, M., Chen, C. M., West, M., Nevins, J. R., and Huang, A. T. (2003), "Gene Expression Predictors of Breast Cancer Outcomes," *The Lancet*, 361, 1590–1596.

Huang, E., West, M., and Nevins, J. R. (2002), "Gene Expression Profiles and Predicting Clinical Characteristics of Breast Cancer," *Hormone Research*, 58, 55–73.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003a), "Summaries of Affymetrix GeneChip Probe-Level Data," *Nucleic Acids Research*, 31, e15.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003b), "Exploration, Normalization, and Summaries of High-Density Oligonucleotide Array Probe-Level Data," *Biostatistics*, 2, 249–264.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005), "Experiments in Stochastic Computation for High-Dimensional Graphical Models," *Statistical Science*, 20, 388–400.

Madigan, D., and York, J. (1995), "Bayesian Graphical Models for Discrete Data," *International Statistical Review*, 63, 215–232.

Nevins, J. R., Huang, E. S., Dressman, H., Pittman, J., Huang, A. T., and West, M. (2003), "Towards Integrated Clinico-Genomic Models for Personalized Medicine: Combining Gene Expression Signatures and Clinical Factors in Breast Cancer Outcomes Prediction," *Human Molecular Genetics*, 12, 153–157.

Pittman, J., Huang, E., Dressman, H., Horng, C. F., Cheng, S. H., Tsou, M. H., Chen, C. M., Bild, A., Iversen, E. S., Huang, A. T., Nevins, J. R., and West, M. (2004), "Integrated Modeling of Clinical and Gene Expression Information for Personalized Prediction of Disease Outcomes," *Proceedings of the National Academy of Sciences*, 101, 8431–8436.

Raftery, A. E. (1995), "Bayesian Model Selection in Social Research," *Sociological Methodology*, 25, 111–163.

Raftery, A. E., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 1197–1208.

Rich, J. N., Hans, C., Jones, B., Iversen, E. S., McClendon, R. E., Rasheed, B. K. A., Dobra, A., Dressman, H. K., Bigner, D. D., Nevins, J. R., and West, M. (2005), "Gene Expression Profiling and Genetic Markers in Glioblastoma Survival," *Cancer Research*, 65, 4051–4058.

Smith, M., and Kohn, R. (1996), "Nonparametric Regression Using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–343.

——— (1997), "A Bayesian Approach to Nonparametric Bivariate Regression," *Journal of the American Statistical Association*, 92, 1522–1535.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005), "Bayesian Variable Selection in Clustering High-Dimensional Data," *Journal of the American Statistical Association*, 100, 602–617.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Marks, J. R., and Nevins, J. R. (2001), "Predicting the Clinical Status of Human Breast Cancer Utilizing Gene Expression Profiles," *Proceedings of the National Academy of Sciences*, 98, 11462–11467.

Wong, F., Carter, C. K., and Kohn, R. (2003), "Efficient Estimation of Covariance Selection Models," *Biometrika*, 90, 809–830.