STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:2004–2021
Published online 2 November 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3118

A false-discovery-rate-based loss framework for selection of interactions

Wei Chen[1,*,†], Debashis Ghosh[2], Trivellore E. Raghunathan[2] and Daniel J. Sargent[3]

[1]*Karmanos Cancer Institute, 716 Harper Professional Building, 4160 John R, Detroit, MI 48201, U.S.A.*
[2]*Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, U.S.A.*
[3]*Division of Biostatistics, Mayo Clinic, Rochester, MN 55905, U.S.A.*

SUMMARY

Interaction effects have been consistently found important in explaining the variation in outcomes in many scientific research fields. Yet, in practice, variable selection including interactions is complicated due to the limited sample size, conflicting philosophies regarding model interpretability, and accompanying amplified multiple-testing problems. The lack of statistically sound algorithms for automatic variable selection with interactions has discouraged activities in exploring important interaction effects. In this article, we investigated issues of selecting interactions from three aspects: (1) What is the model space to be searched? (2) How is the hypothesis-testing performed? (3) How to address the multiple-testing issue? We propose loss functions and corresponding decision rules that control FDR in a Bayesian context. Properties of the decision rules are discussed and their performance in terms of power and FDR is compared through simulations. Methods are illustrated on data from a colorectal cancer study assessing the chemotherapy treatments and data from a diffuse large-B-cell lymphoma study assessing the prognostic effect of gene expressions. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: Bayesian loss; FDR; model building; interaction; Markov chain

1. INTRODUCTION

In the fields of biomedical science, genetics, and epidemiology, complex joint effects among predictors have been consistently found important in explaining variations in outcomes, e.g. the interaction between drugs and bio-markers, between demographic status and risk factors, and

[*] Correspondence to: Wei Chen, Karmanos Cancer Institute, 716 Harper Professional Building, 4160 John R, Detroit, MI 48201, U.S.A.
[†] E-mail: chenw@karmanos.org

Contract/grant sponsor: National Institutes of Health through Karmanos Cancer Institute; contract/grant number: 5 P30 CA022453-25
Contract/grant sponsor: National Institutes of Health through the University of Michigan's Cancer Center; contract/grant number: 5 P30 CA46592

Received 22 December 2006
Accepted 21 September 2007

between genetic variants and environmental factors [1–6]. Identifying interactions is therefore critical in the use of statistical modeling to study complex traits or diseases.

In practice, variable selection including interactions is complicated due to the limited sample size, conflicting philosophies regarding model interpretability, and accompanying amplified multiple-testing problems. In this article, we investigate issues of statistical modeling raised from selecting interactions in the multiple regression setting from three aspects: (1) What is the model space to be searched? (2) How is the hypothesis-testing performed? (3) How to address the multiple-testing issue? We limit our discussion to finding two-way interactions, although higher-level interactions or polynomial terms can be tackled in the same fashion.

Different approaches for variable selection in multiple regressions, when the candidate variables include derived interaction terms (product terms), can be categorized into three classes regarding aspects (1) and (2). The first class of approaches (we name it NC for non-constrained) treats interactions and main effects interchangeably. In this class, no distinction is made between the main effects and the related interactions during the variable selection procedure. NC is commonly implemented in many automatic selection algorithms (e.g. forward, backward, stepwise selection) in practice. The entire model space is searched under this class of methods. However, the intrinsic hypothesis-testing of important regressors suffers from the lack of interpretability of selected models [7–10]. The interactions are interpreted as deviations from additive effects; therefore, testing an additive effect (main effect) of a predictor given its deviation (interaction) in the model is not interpretable.

The second class of approaches excludes interactions whenever their corresponding main effects are not selected. We name it LH here as the inclusion of higher-order terms is constrained by the presence or absence of lower-order terms. Some simple two-stage variable selection methods belong to this class. These methods start with selecting significant main effects, followed by selecting interactions between those selected main effects. A more sophisticated Bayesian hierarchical model, proposed by Chipman [11] imposing the ‘strong heredity rule’ (in their terms), also belongs to this class. The LH methods honor the convention for model interpretability that both interactions and corresponding main effects must be in the model. However, the LH method explores only a portion of the model space, which depends on the results of selecting the main effects.

The final class of approaches automatically forces main effects into the model if the interactions are selected, proposed by Chen [12]. We name it HL here as lower-order terms are imposed by higher-order terms. A sketch of model specification and implementation follows shortly in the next section. From a hypothesis-testing point of view, the HL method includes or excludes a main effect through testing the joint importance of that main effect and all interactions involving that main effect. This approach maintains model interpretability and searches the entire model space.

Despite the divergence among the three classes of variable selection approaches discussed above, all suffer from a multiple hypothesis-testing issue. The problem is amplified if all two-way interactions are explored. Solutions to the multiple-testing issue involve selecting a threshold significance level for a variable to be included in the model in an attempt to control the model Type I error rate. Traditional approaches have aimed at controlling the familywise error rate (FWER), such as Bonferroni and related methods. A more liberal criterion involves the false discovery rate (FDR), proposed by Benjamini and Hochberg [13]. The FDR controls the expected proportion of errors among the rejected hypotheses and generally leads to greater power for detecting alternative hypotheses.

Recently, Ghosh *et al.* [14] established the connection between the FDR and the variable selection problem in the multiple regression setting. It was shown that the FDR is a by-product of a

DOI: 10.1002/sim

hierarchical Bayesian model and the procedures that select variables based on controlling the FDR will have risk optimality properties. Through their work, a new motivation for FDR-controlled procedures for variable selection with interactions is provided.

Extending the work of Ghosh *et al.* [14], we propose automatic variable selection approaches by controlling the posterior expected FDR, a quantity introduced by Genovese *et al.* [15]. Loss functions and corresponding decision rules are proposed with respect to three different approaches (NC, LH, and HL) for handling models with interactions. While related loss functions based on the posterior expected FDR can be found in Müller *et al.* [16], we extend their work to address the issue of variable selection in a multiple regression setting and with interactions. Properties of the decision rules are discussed and their performance in terms of power and FDR is compared through simulations. Additionally, we provide recommendations on choosing the decision rules in different contexts.

In Section 2, we formalize the loss function framework and derive the FDR-controlling decision rules. The performance of the proposed selection rules is studied through simulations in Section 3. In Sections 4 and 5 we illustrate the proposed decision rules on two real data sets. A short discussion is given in Section 6.

2. LOSS FUNCTIONS AND DECISION RULES

We first define appropriate loss functions for each of the approaches (NC, LH, and HL). Suppose that we have a set of covariates including main effects and the corresponding two-way interactions. It is straightforward to extend our method to any higher-order relationships. For the ith covariate, let $d_i \equiv d_i(y) \in \{0,1\}$ denote the decision (given data y, select the ith covariate when $d_i = 1$ and exclude otherwise) and $\gamma_i \in \{0,1\}$ be the truth (unobserved indicator of whether the ith covariate is a true predictor). Let s_1 denote the set of all main effects and s_2 the set of all two-way interactions. For the three approaches for treating the relationship among the regressors, a generic constrained additive loss function can be defined (a more specific definition follows shortly). Let L_{NC} indicate the loss function for the first approach which imposes no constraint, L_{LH} for the second approach which selects lower-order terms first and imposes a constraint on higher-order terms, and L_{HL} for the third approach which selects higher-order terms first and imposes a constraint on lower-order terms. We have

$$L_{\text{NC}} = \sum_{i \in (s_1, s_2)} L(\gamma_i, d_i) \tag{1}$$

$$L_{\text{LH}} = \sum_{i \in s_1} L(\gamma_i, d_i) + \sum_{i \in s_2} L(\gamma_i, d_i^*) \tag{2}$$

where

$$\begin{aligned} d_i^* &= \begin{cases} d_i & \text{if } \textit{all} \text{ of its lower-order terms are selected} \\ 0 & \text{otherwise} \end{cases} \\ L_{\text{HL}} &= \sum_{i \in s_1} L(\gamma_i, d_i^*) + \sum_{i \in s_2} L(\gamma_i, d_i) \end{aligned} \tag{3}$$

DOI: 10.1002/sim

where

$$d_i^* = \begin{cases} d_i & \text{if } \textit{none} \text{ of its higher-order terms are selected} \\ 1 & \text{otherwise} \end{cases}$$

Remark 1

The constraint on d_i^* in (2) corresponds to the strong heredity principle in Chipman [11]. One can relax the constraint to $d_i^* = d_i$ when at least one of the main effects is selected, and $d_i^* = 0$ when none is selected. This modification corresponds to the weak heredity principle from Chipman [11].

2.1. Decision rule for L_{NC}

In general, there are two types of errors in the variable selection problem: selecting a variable that in truth is not a predictor (false discovery) and not selecting a variable that in truth is a predictor (false negative). These two errors can be quantified by two complementary Bayesian losses: posterior expected proportion of discoveries that are false discoveries ($\overline{\text{FDR}}$) and posterior expected proportion of negatives that are false negatives ($\overline{\text{FNR}}$). Let $v_i = P(\gamma_i = 1|y)$ be the marginal posterior probability for the ith regressor, $D = \sum d_i$, and m the total number of regressors in consideration. Since d_i is a function of y, $\overline{\text{FDR}}$ and $\overline{\text{FNR}}$ can be denoted as follows:

$$\overline{\text{FDR}} = \begin{cases} E_{\gamma|y}\left(\left. \dfrac{\sum d_i(1-\gamma_i)}{D} \right| y \right) = \dfrac{\sum d_i(1-v_i)}{D} & \text{if } D>0 \\ 0 & \text{if } D=0 \end{cases}$$

and

$$\overline{\text{FNR}} = \begin{cases} E_{\gamma|y}\left(\left. \dfrac{\sum (1-d_i)\gamma_i}{m-D} \right| y \right) = \dfrac{\sum (1-d_i)v_i}{m-D} & \text{if } D<m \\ 0 & \text{if } D=m \end{cases}$$

One form of Bayesian loss for the NC approach can be defined two dimensionally using the two complementary losses $\overline{\text{FDR}}$ and $\overline{\text{FNR}}$ ($E(L_{\text{NC}}|y) = \{\overline{\text{FDR}}, \overline{\text{FNR}}\}$). Controlling one dimension and minimizing the other is a straightforward approach for minimizing a two-dimensional loss [17]. Determining whether to control $\overline{\text{FDR}}$ or $\overline{\text{FNR}}$ depends on the objective of the analysis. For example, if the objective is to restrain the selection of variables that in truth are not in the model, one would like to control the $\overline{\text{FDR}}$ at a certain level while minimizing the $\overline{\text{FNR}}$. For simplicity, here and afterwards we discuss minimization of $\overline{\text{FNR}}$ subject to controlling $\overline{\text{FDR}}$ at a significance level α. Minimizing $\overline{\text{FDR}}$ subject to $\overline{\text{FNR}} \leqslant \alpha$ can be easily derived in a similar manner. The choice of α, which is different from a significance level for the conventional familywise error rate, is discussed in more detail in the simulation studies.

To minimize $E(L_{\text{NC}}|y)$, one can find a set of thresholds $\{t\}$ such that a decision $d_i \equiv I(v_i > t)$, $i = 1, \ldots, m$, results in $\overline{\text{FDR}} \leqslant \alpha$. Since $\overline{\text{FNR}}$ is minimized by $\min\{t\}$, the optimal threshold

$$t_{\text{NC}}^* \equiv \min\{t : \overline{\text{FDR}} \leqslant \alpha\}$$

minimizes $E(L_{\text{NC}}|y)$. The proof follows directly from Müller *et al.* [16]. Note that the optimal rule has the same form as in Müller *et al.* However, in their work only main effects were considered individually in calculating sample size for gene expression experiments. Variable selection with

DOI: 10.1002/sim

interactions in a multiple regression setting was not addressed. Note that Devlin *et al.* [3] proposed a model selection procedure for finding epistatic loci in genetic association studies using the FDR. Within the framework assumed in the decision rule here, their procedure satisfies this optimality property.

Remark 2

By the law of iterated expectations, control of the $\overline{\text{FDR}}$ equals control of the FDR if we average with respect to the distribution of the data. Since one can simulate the entire posterior distribution of the FDR, that would provide information on the actual FDR. We estimated the actual mean of FDR in the simulation studies.

2.2. *Decision rule for* L_{LH}

As the loss function L_{LH} defined in (2) is additive, we define the losses $\overline{\text{FDR}}$ and $\overline{\text{FNR}}$ within the set of main effects (s_1) and interactions (s_2) separately using a subscript to distinguish them. To minimize the posterior expected loss ($E(L_{\text{LH}}|y)$) given a pre-specified control level α_{L} for main effects and α_{H} for interactions, we have

$$\begin{aligned}\min\{E(L_{\text{LH}}|y)\} &= \min\{E(L_{\text{LH}}|y)_{s_1}, E(L_{\text{LH}}|y)_{s_2}\}\\ &= \min\{\overline{\text{FNR}}_{s_1}, \overline{\text{FNR}}_{s_2}\,|\,\overline{\text{FDR}}_{s_1}\leqslant\alpha_{\text{L}}, \overline{\text{FDR}}_{s_2}\leqslant\alpha_{\text{H}}\}\end{aligned} \tag{4}$$

We can minimize $E(L_{\text{LH}}|y)$ sequentially using the additive nature of the loss function. As the decision in s_2 is controlled by the decision of lower-order terms in s_1, we start by minimizing the posterior expected loss in s_1 ($E(L_{\text{LH}}|y)_{s_1}$), followed by minimizing the posterior expected loss in s_2 ($E(L_{\text{LH}}|y)_{s_2}$). After a decision is made for terms in s_1, a subset (denoted as s_2') of the higher-order terms in s_2 are excluded from consideration due to the constraint. Hence, those terms are not involved in making the decision to minimize the posterior expected loss in s_2. A decision will be made for the remaining terms in the complement of s_2' (denoted as $\overline{s_2'}$). In total, two decisions (one for terms in s_1 and the other for terms in $\overline{s_2'}$) are required. Algorithm 1 illustrates the steps to construct the two decision rules.

Algorithm 1

Decision rules for L_{LH}

1. Find an optimal threshold $t^*_{\text{LH}_1}$, such that the decision $d_i = I(v_i > t^*_{\text{LH}_1}), i\in s_1$, minimizes $E(L_{\text{LH}}|y)_{s_1}$. We have

$$t^*_{\text{LH}_1} = \min\{t : \overline{\text{FDR}}_{s_1} \leqslant \alpha_{\text{L}}\}$$

and

$$\overline{\text{FDR}}_{s_1} = \begin{cases} \dfrac{\sum_{i\in s_1} d_i(1-v_i)}{\sum_{i\in s_1} d_i} & \text{if } \sum_{i\in s_1} d_i > 0 \\ 0 & \text{if } \sum_{i\in s_1} d_i = 0 \end{cases}$$

2. Within s_2, identify a subset s_2' whose corresponding lower-order terms are *NOT* selected in step 1. Set $d_i = 0, i \in s_2'$.

DOI: 10.1002/sim

3. Find an optimal threshold $t^*_{\mathrm{LH}_2}$, such that the decision $d_i = I(v_i > t^*_{\mathrm{LH}_2}), i \in \overline{s_2'}$, minimizes $E(L_{\mathrm{LH}}|y)_{s_2}$. We have

$$t^*_{\mathrm{LH}_2} = \min\{t : \overline{\mathrm{FDR}}_{\overline{s_2'}} \leqslant \alpha_{\mathrm{H}}\}$$

and

$$\overline{\mathrm{FDR}}_{\overline{s_2'}} = \begin{cases} \dfrac{\sum_{i\in\overline{s_2'}} d_i(1-v_i)}{\sum_{i\in\overline{s_2'}} d_i} & \text{if } \sum_{i\in\overline{s_2'}} d_i > 0 \\ 0 & \text{if } \sum_{i\in\overline{s_2'}} d_i = 0 \end{cases}$$

This sequential decision rule can be easily implemented in a Bayesian variable selection algorithm. The selected variables will ensure a ‘well-formulated’ model in the sense of Peixoto [9].

Theorem 1
The posterior expected total FDR ($\overline{\mathrm{FDR}}_T$) under the loss function L_{LH} is controlled at max $(\alpha_{\mathrm{L}}, \alpha_{\mathrm{H}})$.

Proof
By definition,

$$\overline{\mathrm{FDR}}_T = \begin{cases} \dfrac{\sum_{i\in s_1} d_i(1-v_i) + \sum_{i\in\overline{s_2'}} d_i(1-v_i)}{D_T} & D_T > 0 \\ 0 & D_T = 0 \end{cases}$$

$$= \begin{cases} \dfrac{D_{s_1}\overline{\mathrm{FDR}}_{s_1} + D_{\overline{s_2'}}\overline{\mathrm{FDR}}_{\overline{s_2'}}}{D_{s_1} + D_{\overline{s_2'}}} & D_T > 0 \\ 0 & D_T = 0 \end{cases}$$

where $D_T = \sum_{i\in s_1} d_i + \sum_{i\in\overline{s_2'}} d_i$, $D_{s_1} = \sum_{i\in s_1} d_i$, and $D_{\overline{s_2'}} = \sum_{i\in\overline{s_2'}} d_i$. The $\overline{\mathrm{FDR}}_T$ is a weighted average of $\overline{\mathrm{FDR}}_{s_1}$ and $\overline{\mathrm{FDR}}_{\overline{s_2'}}$. Since $\overline{\mathrm{FDR}}_{s_1} \leqslant \alpha_{\mathrm{L}}$ and $\overline{\mathrm{FDR}}_{\overline{s_2'}} \leqslant \alpha_{\mathrm{H}}$, we have $\overline{\mathrm{FDR}}_T \leqslant \max(\alpha_{\mathrm{L}}, \alpha_{\mathrm{H}})$. □

2.3. Decision rule for L_{HL}

Similar to the rule for L_{LH}, we can minimize the posterior expected loss ($E(L_{\mathrm{HL}}|y)$) sequentially in reversed order. Since the mandatory inclusion of terms in s_1 is controlled by the decisions of their higher-order terms in s_2, we start with minimizing the posterior expected loss in s_2 ($E(L_{\mathrm{HL}}|y)_{s_2}$) followed by minimizing the posterior expected loss in s_1 ($E(L_{\mathrm{HL}}|y)_{s_1}$). After a decision is made in s_2, a subset (denoted as s_1') of the lower-order terms in s_1 are forced to be included in the model due to the constraint. Hence, those terms are not involved in the decision to minimize $E(L_{\mathrm{HL}}|y)_{s_1}$. A decision is required for the terms in the complement of s_1' (denoted by $\overline{s_1'}$). A similar procedure to define the decision rules in this case is provided in Algorithm 2.

DOI: 10.1002/sim

Algorithm 2
Decision rules for L_{HL}

1. Find an optimal threshold $t^*_{\mathrm{HL}_1}$, such that the decision $d_i = I(v_i > t^*_{\mathrm{HL}_1}), i \in s_2$, minimizes $E(L_{\mathrm{HL}}|y)_{s_2}$. We have

$$t^*_{\mathrm{HL}_1} = \min\{t : \overline{\mathrm{FDR}}_{s_2} \leqslant \alpha_{\mathrm{H}}\}$$

and

$$\overline{\mathrm{FDR}}_{s_2} = \begin{cases} \dfrac{\sum_{i\in s_2} d_i(1-v_i)}{\sum_{i\in s_2} d_i} & \text{if } \sum_{i\in s_2} d_i > 0 \\ 0 & \text{if } \sum_{i\in s_2} d_i = 0 \end{cases}$$

2. Within s_1, identify a subset s'_1 whose corresponding higher-order terms are selected in step 1. Set $d_i = 1, i \in s'_1$. Thus, we have $\overline{\mathrm{FDR}}_{s'_1} = \sum_{i\in s'_1}(1-v_i)/\sum_i I(i \in s'_1)$.
3. Find an optimal threshold $t^*_{\mathrm{HL}_2}$, such that the decision $d_i = I(v_i > t^*_{\mathrm{HL}_2}), i \in \overline{s'_1}$, minimizes $E(L_{\mathrm{HL}}|y)_{s_1}$. We have

$$t^*_{\mathrm{LH}_2} = \min\{t : \overline{\mathrm{FDR}}_{\overline{s'_1}} \leqslant \alpha_{\mathrm{L}}\}$$

and

$$\overline{\mathrm{FDR}}_{\overline{s'_1}} = \begin{cases} \dfrac{\sum_{i\in \overline{s'_1}} d_i(1-v_i)}{\sum_{i\in \overline{s'_1}} d_i} & \text{if } \sum_{i\in \overline{s'_1}} d_i > 0 \\ 0 & \text{if } \sum_{i\in \overline{s'_1}} d_i = 0 \end{cases}$$

Algorithm 2 guarantees that the selected variables constitute a 'well-formulated' model; however compared with Algorithm 1, Algorithm 2 increases the chance of detecting the true interactions due to a non-restrained searching space.

Lemma 1
The posterior expected total FDR ($\overline{\mathrm{FDR}}_T$) under the loss function L_{HL} is controlled at $\max(\alpha_{\mathrm{L}}, \alpha_{\mathrm{H}})$.

Proof

$$\overline{\mathrm{FDR}}_T = \begin{cases} \dfrac{\sum_{i\in s'_1}(1-v_i) + \sum_{i\in \overline{s'_1}} d_i(1-v_i) + \sum_{i\in s_2} d_i(1-v_i)}{D_T} & D_T > 0 \\ 0 & D_T = 0 \end{cases}$$

$$= \begin{cases} \dfrac{D_{s'_1}\overline{\mathrm{FDR}}_{s'_1} + D_{\overline{s'_1}}\overline{\mathrm{FDR}}_{\overline{s'_1}} + D_{s_2}\overline{\mathrm{FDR}}_{s_2}}{D_{s'_1} + D_{\overline{s'_1}} + D_{s_2}} & D_T > 0 \\ 0 & D_T = 0 \end{cases}$$

DOI: 10.1002/sim

where $D_{s_1'}=\sum_i I(i\in s_1')$, $D_{\overline{s_1'}}=\sum_{i\in\overline{s_1'}} d_i$, $D_{s_2}=\sum_{i\in s_2} d_i$, and $D_{s_1'}+D_{\overline{s_1'}}+D_{s_2}=D_T$. We have $\overline{\mathrm{FDR}}_{\overline{s_1'}}\leqslant\alpha_{\mathrm{L}}$ and $\overline{\mathrm{FDR}}_{s_2}\leqslant\alpha_{\mathrm{H}}$. All the elements in s_1' are the main effects whose interactions are selected. Under the constraint that main effects must be selected if interactions are selected, the marginal posterior probability of a main effect is always equal to or greater than that of its interactions. Thus, $\overline{\mathrm{FDR}}_{s_1'}$ is also controlled at α_{H}. Since $\overline{\mathrm{FDR}}_T$ is a weighted average of $\overline{\mathrm{FDR}}_{s_1'}$, $\overline{\mathrm{FDR}}_{\overline{s_1'}}$, and $\overline{\mathrm{FDR}}_{s_2}$, we conclude that $\overline{\mathrm{FDR}}_T\leqslant\max(\alpha_{\mathrm{L}},\alpha_{\mathrm{H}})$. □

Remark 3

When different control levels α_{L} for s_1 and α_{H} for s_2 are desired in L_{NC}, similar sequential decision rules can be derived to minimize the posterior expected loss function $E(L_{\mathrm{NC}}|y)$. The steps are similar to Algorithm 1 or 2 except that the order makes no difference and there is no constraint on either main effects or interactions. This provides flexibility of controlling $\overline{\mathrm{FDR}}$ under L_{NC}.

Remark 4

The proposed decision rule for the NC approach is a single-step-testing procedure. The decision rules for LH and HL approaches are two-step procedures. Regressors are divided into two blocks: main effects and interaction terms. The threshold t^* for decision $d_i=I(v_i>t^*)$ within each block is an optimal threshold. Even though a step-wise procedure $d_i=I(v_i>t_i)$ could be entertained, it is non-trivial to identify an optimal threshold in this context.

2.4. Applications to the linear regression model

The three proposed loss-function-based selection rules can be easily incorporated into a general Bayesian hierarchical regression model. A stochastic search variable selection (SSVS) algorithm by George *et al.* [18] will be employed here with a slight modification. For a simple linear regression with normal errors,

$$Y=X\beta+\varepsilon \tag{5}$$

a mixture normal prior is specified for each coefficient β_i,

$$\beta_i|\gamma_i\sim(1-\gamma_i)\mathrm{N}(0,\tau^2)+\gamma_i\mathrm{N}(0,c^2\tau^2)$$

The binary latent variables γ_i are incorporated in the model. If $\gamma_i=1$, the ith regressor is a true predictor. According to George *et al.* [18], when $\gamma_i=0$, β_i is closely centered at zero with a small variance τ^2. When $\gamma_i=1$, β_i is allowed to have large positive or negative effects using a large value of c. c can be interpreted as the prior odds that ith regressor should be excluded when β_i is very close to zero. Here we followed the recommendation of choosing c and τ given in George *et al.* [18].

We assume an Inverse Gamma prior for the residual variance σ^2, a Bernoulli prior for $p_i=\Pr(\gamma_i=1)$, and a Beta hyperprior for p_i:

$$\sigma^2\sim\mathrm{IG}(v/2,v\lambda/2),\quad \gamma_i\sim\mathrm{Bern}(p_i)\quad\text{and}\quad p_i\sim\mathrm{Beta}(a,b) \tag{6}$$

where $a=(\mathrm{mean}-2\times\mathrm{mode}\times\mathrm{mean})/(\mathrm{mean}-\mathrm{mode})$, $b=(1-2\times\mathrm{mode})\times(1-\mathrm{mean})/(\mathrm{mean}-\mathrm{mode})$.

The shape of the Beta distribution shall be skewed toward small values, suggesting that *a priori* the number of true predictors is just a small fraction of number of all the regressors. The mean

DOI: 10.1002/sim

for the Beta distribution is the average prior probability to be a true predictor and the distance between mean and mode represents uncertainties about the number of the true predictors.

The Bernoulli prior for γ_i in (6) indicates that no constraint is forced among related predictors. This is the prior specification for the loss function L_{NC}. Two different degenerate Bernoulli distributions were used for γ_i according to the two loss functions L_{LH} and L_{HL}.

For the prior based on L_{LH}:

$$\begin{aligned}&\text{For } i\in s_1 \quad \gamma_i\sim \mathrm{Bern}(p_i)\\ &\text{For } i\in s_2 \quad \begin{cases}\gamma_i\sim\mathrm{Bern}(p_i) & \text{if } \textit{all} \text{ of its lower-order terms are selected}\\ \gamma_i\sim\mathrm{Bern}(0) & \text{o.w.}\end{cases}\end{aligned} \tag{7}$$

For the prior based on L_{HL}:

$$\begin{aligned}&\text{For } i\in s_1 \quad \begin{cases}\gamma_i\sim\mathrm{Bern}(p_i) & \text{if } \textit{none} \text{ of its higher-order terms are selected}\\ \gamma_i\sim\mathrm{Bern}(1) & \text{o.w.}\end{cases}\\ &\text{For } i\in s_2 \quad \gamma_i\sim \mathrm{Bern}(p_i)\end{aligned} \tag{8}$$

The marginal posterior distribution of γ_i is our primary interest for variable selection. This posterior is estimated by the Gibbs sampling method through iteratively updating the parameters from their full conditional distributions.

3. SIMULATION

Monte Carlo simulations were employed to assess the performance of the three proposed decision rules concerning the total FDR and the power of detecting true regressors. The sensitivity of the prior distribution, sample size, and different control values $(\alpha_{\mathrm{L}}, \alpha_{\mathrm{H}})$ were also studied.

We considered variable selection on $p=55$ (10 main effects and 45 two-way interactions) with three different sample sizes $n=15, 30$, and 100. The true model was $Y=0.3+x_1+x_2+1.5x_3+1.5x_4+x_1x_2+x_1x_4+\varepsilon$. For simplicity, the main effects $\{x_i\}, i=1,\ldots,10$, were independent and identically distributed N(0, 1). The variance of ε was set to 1. We set $c=50$, $\tau=0.02$, and the (mode, mean) for the Beta hyperprior at (0.05,0.06). For each sample size, 100 simulated replications were performed.

For each of the three loss functions (L_{NC}, L_{LH}, and L_{HL}), various control values $(\alpha_{\mathrm{L}}, \alpha_{\mathrm{H}})$ were used to identify the true predictors. Under each set of $(\alpha_{\mathrm{L}}, \alpha_{\mathrm{H}})$, the total FDR (proportion of falsely discovered predictors among discovered predictors) and power (proportion of discovered true predictors among the true predictors) were calculated at each simulation and averaged across 100 replications. Table I summarizes the results of the simulations.

Overall, the total FDR is well controlled at significance levels $(\alpha_{\mathrm{L}}, \alpha_{\mathrm{H}})$ under all three decision rules (Table I). When $n<p$, the L_{HL} criterion has higher power than the other two criteria. The difference in power among the three criteria decreases when the sample size increases. The three loss functions achieved the same power for detecting true predictors when $n=100$, showing that the impact of the prior decreases as the sample size increases in the Bayesian framework.

To examine the sensitivity to the prior, we repeated our algorithm for several choices of c (10, 50, and 100) and τ (0.02 and 0.05). The simulation results changed very little for values of c

Table I. Simulation results with mode$=0.05$, mean$=0.06$, and $\mathrm{Var}(\varepsilon)=1$.

		(α_L, α_H)																	
		(0.05, 0.05)			(0.05, 0.2)			(0.05, 0.8)			(0.2, 0.05)			(0.2, 0.2)			(0.2, 0.8)		
n		NC	LH	HL	NC	LH	HL	NC	LH	HL	NC	LH	HL	NC	LH	HL	NC	LH	HL
15	FDR	0.04	0	0.02	0.06	0	0.03	0.61	0	0.53	0.04	0	0.07	0.06	0	0.08	0.60	0	0.53
	Power	0.03	0.02	0.11	0.03	0.02	0.12	0.11	0.02	0.62	0.04	0.05	0.24	0.05	0.05	0.24	0.13	0.05	0.62
30	FDR	0.01	0	0.01	0.05	0	0.01	0.58	0.13	0.65	0.02	0.01	0.04	0.06	0.01	0.03	0.53	0.14	0.65
	Power	0.20	0.24	0.74	0.25	0.26	0.81	0.41	0.31	0.99	0.34	0.33	0.79	0.38	0.36	0.85	0.54	0.43	0.99
100	FDR	0	0	0	0	0	0	0.57	0.57	0.68	0.14	0.14	0	0.14	0.14	0	0.60	0.60	0.68
	Power	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Total number of regressors $p=55$.
The total FDR and power are averaged over 100 simulations.

Table II. Simulation results with mode$=0.1$, mean$=0.15$, and $\mathrm{Var}(\varepsilon)=1$.

		(α_L, α_H)																	
		(0.05, 0.05)			(0.05, 0.2)			(0.05, 0.55)			(0.2, 0.05)			(0.2, 0.2)			(0.2, 0.55)		
n		NC	LH	HL	NC	LH	HL	NC	LH	HL	NC	LH	HL	NC	LH	HL	NC	LH	HL
15	FDR	0.11	0	0.10	0.27	0	0.12	0.68	0	0.45	0.11	0.02	0.30	0.27	0.02	0.31	0.65	0.03	0.46
	Power	0.11	0.06	0.30	0.12	0.06	0.31	0.18	0.06	0.53	0.15	0.10	0.48	0.16	0.10	0.48	0.22	0.11	0.58
30	FDR	0.01	0	0.01	0.03	0.01	0.03	0.34	0.12	0.30	0.04	0.06	0.06	0.06	0.06	0.04	0.30	0.15	0.30
	Power	0.36	0.53	0.80	0.41	0.58	0.86	0.53	0.65	0.95	0.52	0.62	0.83	0.58	0.67	0.89	0.69	0.76	0.96
100	FDR	0	0	0	0.01	0	0	0.27	0.25	0.27	0.14	0.14	0	0.15	0.14	0	0.35	0.33	0.27
	Power	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

and τ in this range (table not shown). As the setting of mean for the Beta hyperprior represents a prior belief of average probability of being a true predictor, we changed the setting of (mode, mean) from $(0.05, 0.06)$ to $(0.1, 0.15)$ and re-ran our algorithm. In the new setting, the basic shape of the Beta distribution was not changed, except that the new prior assumed *a priori* more true predictors among the 55 regressors. The power and FDR were unchanged when $n=100$. However, as expected we observed an increase in power and FDR when $n<p$ (Table II). This result indicates that when the sample size is smaller than the number of regressors, the prior is highly informative and the posterior expected losses are affected consequently. We increased the noise in the simulation by increasing $\mathrm{Var}(\varepsilon)$ to 5. In this scenario, the power decreased significantly under the L_{NC} and L_{LH} criteria (Table III).

In general, to achieve greater power when $n<p$, we suggest using the L_{HL} loss function and small (mode, mean) for the Beta prior. If the *a priori* specified number of true predictors (mean $\times$ p) is larger than the sample size (n), it will result in a large number of false positives and such a model is rarely of interest to statisticians or scientists.

Note that changing significance levels (α_L, α_H) does not necessarily change the total FDR, because the decision space is discrete in nature. Furthermore, increasing (α_L, α_H) does not necessarily increase total FDR. If higher (α_L, α_H) results in selecting more regressors which are the

Table III. Simulation results with mode $=0.05$, mean $=0.06$, and $\mathrm{Var}(\varepsilon)=5$.

		(α_L, α_H)																	
		(0.05, 0.05)			(0.05, 0.2)			(0.05, 0.8)			(0.2, 0.05)			(0.2, 0.2)			(0.2, 0.8)		
n		NC	LH	HL	NC	LH	HL	NC	LH	HL	NC	LH	HL	NC	LH	HL	NC	LH	HL
15	FDR	0.04	0	0.02	0.05	0	0.02	0.50	0	0.38	0.04	0	0.06	0.05	0	0.06	0.50	0	0.38
	Power	0.01	0.01	0.04	0.01	0.01	0.04	0.05	0.01	0.32	0.01	0.01	0.08	0.01	0.01	0.08	0.06	0.01	0.32
30	FDR	0	0	0	0.02	0	0.01	0.57	0	0.41	0	0.01	0.01	0.02	0.01	0.02	0.56	0.01	0.41
	Power	0.03	0.01	0.12	0.04	0.01	0.14	0.14	0.01	0.59	0.04	0.05	0.24	0.05	0.05	0.25	0.15	0.05	0.60
100	FDR	0	0	0	0.01	0	0.01	0.45	0.29	0.65	0.01	0.04	0.04	0.02	0.04	0.01	0.43	0.29	0.65
	Power	0.58	0.62	0.81	0.62	0.65	0.9	0.77	0.77	1	0.68	0.69	0.83	0.71	0.72	0.91	0.86	0.84	1

true predictors, the total FDR will decrease due to the unchanged numerator (number of falsely discovered regressors) and increased denominator (number of discovered regressors). For instance, in Table I under the HL approach when $n=30$, the total FDR $=0.03$ at $(\alpha_L=0.2, \alpha_H=0.2)$ is lower than the total FDR $=0.04$ at $(\alpha_L=0.2, \alpha_H=0.05)$. However, this phenomenon apparently does not happen often. In general, the total FDR increases when (α_L, α_H) increase (see Tables I–III).

The significance levels (α_L, α_H) are arbitrary when there is a sufficient number of samples. While high significance levels result in more falsely selected regressors, it would not necessarily cause a worse prediction to the outcome but may lead to a less parsimonious model. When the sample size is limited $(n<p)$, the posterior expected loss will be highly influenced by priors. To increase power, (α_L, α_H) can be set according to the prior specifications. For example, if we assume that on average the probability to be a true predictor is 0.06 (as the mean chosen for the Beta prior in Tables I and III), a choice of $\alpha_H=0.8$ results in an average posterior probability of at least 0.2 among the selected interactions. An average ratio of posterior to prior probability (or the alternative to null hypothesis ratio) for the selected interactions is about threefold, a rule-of-thumb that has been commonly implemented in identifying differential gene expressions in microarray analyses. If we assume that the prior probability to be a true predictor is 0.15 (as in Table II), a 3-fold posterior to prior probability ratio results in a choice of $\alpha_H=0.55$. A similar idea can be applied to choose α_L.

4. COLORECTAL CANCER STUDY

We now apply our proposed decision rules to a phase III clinical trial for the treatment of advanced colorectal cancer initiated by Mayo Clinic in 1997. A total of 1705 patients were enrolled, of which 513 were genotyped for 23 biomarkers. The biomarkers were selected based on previous reports of interaction with the chemotherapies used in the clinical trial. Two experimental treatments 5-fluorouracil + oxaliplatin and oxaliplatin + irinotecan were compared with the standard treatment of 5-fluorouracil + irinotecan. Hereafter, we refer to them as arm F, G, and A, respectively. At the end of the study, the experimental treatment F was approved by the Food and Drug Administration for the treatment of patients with advanced colorectal cancer [19, 20].

A secondary goal of this clinical trial was to determine the role of these biomarkers in predicting the progression-free survival. We are, therefore, interested in exploring all of the two-way

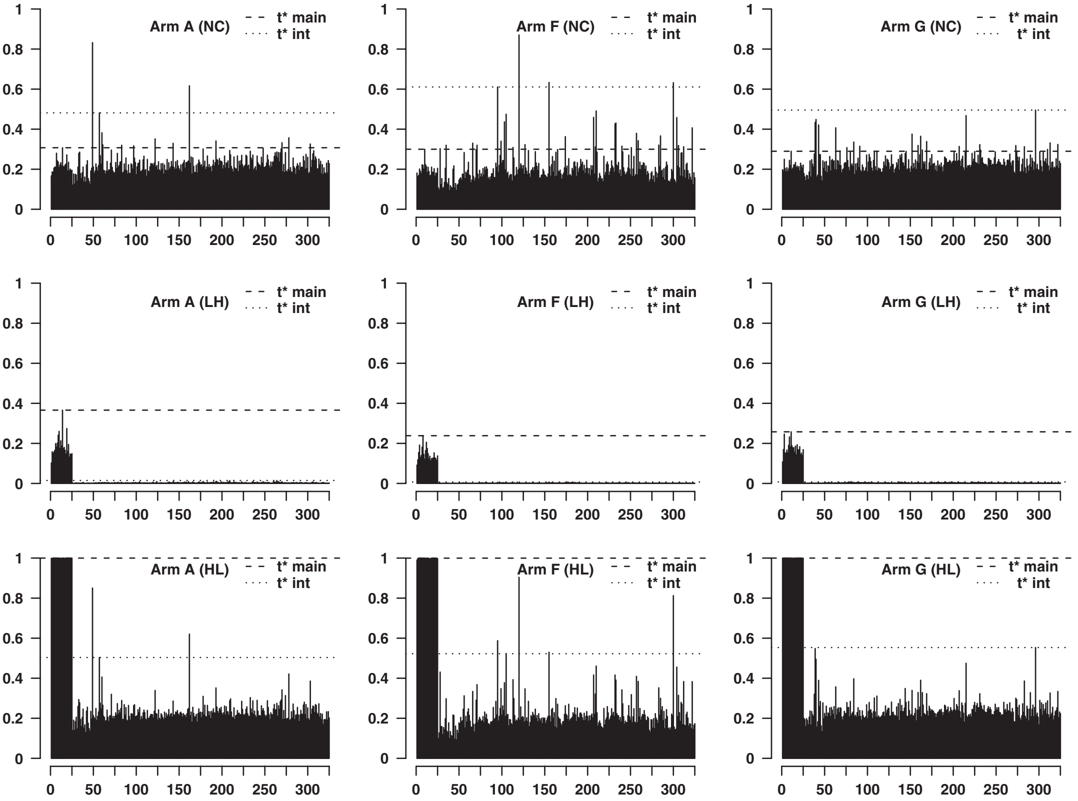

Figure 1. Marginal posterior probabilities in colorectal cancer study. The X-axis is the regressor; the Y-axis is the marginal posterior probability. The first 25 regressors are main effects and the rest are two-way interactions. t^* is the optimal threshold when $\alpha_L = 0.001$ and $\alpha_H = 0.3$.

interactions of the biomarkers and assessing their interactions with different treatment regimens. We dichotomized each of the 23 biomarkers into two categories: mutant and wild type. The total number of main effects is 25 (23 biomarkers plus variables AGE and SEX). Therefore, the number of two-way interactions is 300, which exceeds the sample size under stratified modeling (115 in arm A, 292 in arm F, and 106 in arm G).

The outcome variable in this case is a censored continuous variable; we therefore used the data-augmentation approach [21, 22] to replace the outcome variable Y with a latent complete variable Z in the regression model (5). For each individual j, we have

$$\log(y_j)\begin{cases} = z_j & \text{if } c_j = 1 \\ < z_j & \text{if } c_j = 0 \end{cases}$$

where $c_j, j = 1, \ldots, n$, is the censoring indicator. By assuming normal errors for the regression of log-transformed survival times, we achieved computational efficiency as the full conditional distribution of Z is simply a truncated normal distribution. We could alternatively replace the log-normal model with Weibull, exponential, or Gamma models, which are widely used parametric

Table IV. Est.(Sd.) of coefficients for the selected predictors.

	Arm A ($n=115$)		Arm F ($n=292$)		Arm G ($n=106$)	
Variables	L_{NC}	L_{HL}	L_{NC}	L_{HL}	L_{NC}	L_{HL}
Intercept	5.22(0.04)	4.94(0.09)	5.62(0.06)	5.6(0.18)	—	5.27(0.07)
Age		0.12(0.09)				
abcb1_2677				−0.33(0.26)		
abcb1_3435				−0.08(0.17)		
abcc1_34215		0.66(0.2)		−0.5(0.28)		
abcc2_24				0.16(0.19)		
abcc2_c1515y						−0.39(0.35)
abcc2_v417i				0.81(0.23)		
cyp3a4						0.41(0.33)
dpyd_9a		0.28(0.2)				0.19(0.19)
gstp1_I105v				−0.48(0.13)		
gstp_114		−0.38(0.38)				
tyms_1494del				−0.55(0.14)		
ABCG2Q141K		0.1(0.22)				
Age*ABCG2Q141K	−0.61(0.22)	−0.73(0.21)				
abcb1_2677*abcb1_3435				0.75(0.28)		
abcb1_3435*abcc2_v417i			−0.13(0.11)	−0.98(0.24)		
abcc1_34215*abcc2_24			0.33(0.12)	0.8(0.3)		
abcc1_34215*dpyd_9a	−0.51(0.26)	−1.12(0.35)				
gstp1_I105v*tyms_1494del			0.03(0.11)	0.66(0.2)		

The cell is empty if the variable is not selected.

models for censored survival data. The adaptive rejection sampling algorithm [23] can be utilized for Bayesian inference in this case.

For this data set, we choose $c=10$, $\tau=0.05$, and (mode, mean) $=(0.2, 0.25)$ for a Beta prior with 20 000 iterations of Gibbs sampling. The marginal posterior probabilities for the three modeling approaches (NC, LH, and HL), which were estimated by averaging the event of $\gamma_i=1$ ($i=1,\ldots,325$) across the entire Markov chain Monte Carlo (MCMC) iterations, are shown in Figure 1.

As shown in Figure 1, the loss function L_{LH} resulted in the lowest marginal posterior probabilities. Because of the constraints, all the interaction terms have almost zero probability. Additionally, no important main effects were found under L_{LH} or L_{NC} even when the α_{L} was set at 0.6. Under the loss functions L_{NC} and L_{HL}, several interactions show relatively high probabilities. Exploring all possible two-way interactions under these two loss functions led to the discovery of a complex model which may predict the outcome with greater power. As 300 interaction terms were explored in this case, the marginal posterior probabilities for the main effects are almost equal to one under L_{HL}, because the importance of a variable in the HL approach is determined by the joint importance of its own main effect and all the interaction terms involving that main effect. Also note that the interactions that have high posterior probability in arm A are different from those in arm F (Figure 1), which indicates the existence of treatment–biomarker interactions. This finding is not unexpected, as different biomarkers were selected to be specific to individual treatments.

We set (α_{L}, α_{H}) at (0.001, 0.3) for all the loss functions (no main effects were selected under L_{NC} and L_{LH} even if we raised α_{L} to 0.6). In order to estimate the magnitude of the effect of the selected terms, we ran another MCMC which included only the selected terms. The estimated posterior mean and standard error are shown in Table IV. The selected variables using L_{NC} are a subset of those selected under L_{HL}. Several interactions were identified as significant. For example, under treatment A, older patients with mutant marker ABCG2Q141K tend to have a shorter progression-free survival time using either criteria. External validation of the findings is worth investigating.

We also ran the variable selection procedure using the main effects only model, and no main effect was found significant or to have high marginal posterior probabilities in the multiple regression setting. This result highlights the importance of using the proposed variable selection approaches to explore complex models to improve the understanding of the association of the biomarkers with outcomes of interest.

5. LYMPHOMA STUDY

Our next example is based on gene expression data for patients with diffuse large-B-cell lymphoma (DLBCL). Studies have demonstrated that gene-expression signatures may be useful to predict the prognosis in patients with DLBCL [24–26]. Lossos *et al.* [27] used the quantitative reverse-transcriptase polymerase chain reaction (RT-PCR) to measure the expression of genes from 66 independent DLBCL patients. They studied 36 genes which had been reported to predict survival from past microarray studies. A model to predict the overall survival in DLBCL was proposed using a combination of weighted expressions of six genes (LMO2, BCL6, FN1, CCND2, SCYA3, and BCL2). They concluded that 'measurement of the expression of six genes is sufficient to predict overall survival in diffuse large-B-cell lymphoma.'

We attempted to validate their conclusion by exploring the possible existence of interactions in a multiple regression model. We re-analyzed the lymphoma data with a reduced and a full model. The reduced model was a multiple regression model with all of the 36 genes and only six interactions (p53∗BCL2, p53∗CR2, PRDM1∗IRF4, BCL6∗CCND2, BCL6∗IRF4, BCL6∗SCYA3/CCL3). These six interactions were reported directly or indirectly correlated in different contexts [28–31]. The full model was a multiple regression model with 36 genes and all the two-way interactions.

We began by log transforming and normalizing raw RT-PCR data using the software provided by Eisen Lab at ⟨http://rana.lbl.gov/EisenSoftware.htm⟩. A heatmap with the gene names after two-way hierarchical clustering is shown in Plate 1. Since there was no censoring in this data set, to reduce computational complexity we assumed that the survival followed a normal distribution after log-transformation.

In the reduced model, the number of regressors was $p=36+6=42$, which was smaller than the sample size $n=66$. We set $c\equiv 50$ and $\tau\equiv 0.02$. Two sets of the mode and mean of the Beta hyperprior were used: $(0.2, 0.5)$ and $(0.5, 0.5)$. The marginal posterior probabilities of $\{\gamma_i\}$ were plotted in Figure 2(a) and (b). The marginal posterior probabilities were higher when the prior mean and mode were larger. However, none of the six interactions had posterior probability greater than 0.2 in this data set.

On the basis of an assumption that some interactions may exist but are not yet reported, we proceeded to search all possible interactions using the full model which includes all two-way interactions. In this case the total number of regressors was $p=666$ (36 main effects plus 630 two-way interactions), which was much larger than the sample size $n=66$. The loss function L_{HL} criterion and its corresponding SSVS probability model were used, since a better performance of L_{HL} criterion when $n<p$ was demonstrated in the simulation study (Section 3). However, we

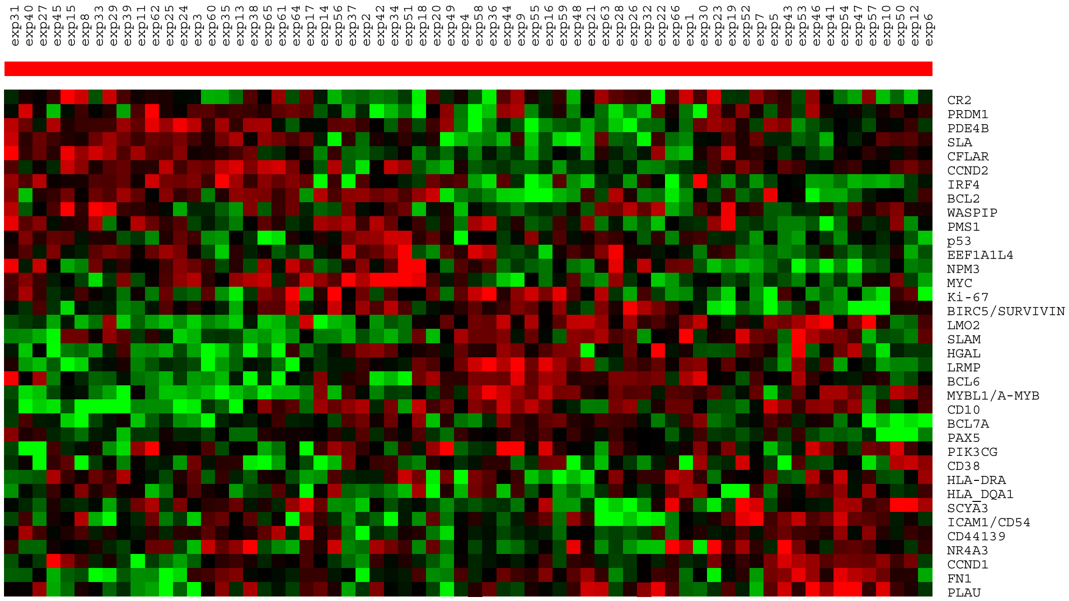

Plate 1. Heatmap of RT-PCR data for diffuse large-B-cell lymphoma patients.

DOI: 10.1002/sim

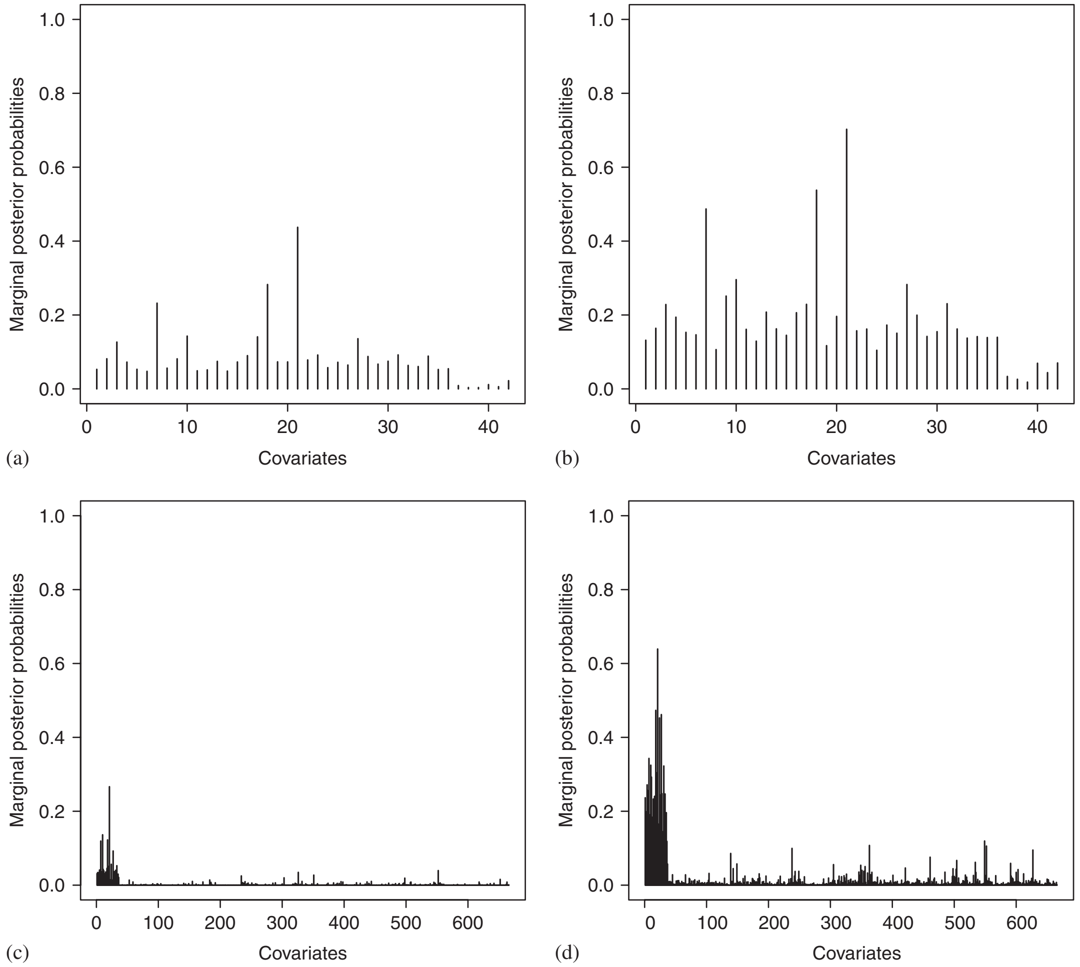

Figure 2. Marginal posterior probabilities in diffuse large-B-cell lymphoma data. The first 36 regressors are the main effects, and the rest are the interactions: (a) Reduced model, mode $=0.2$, mean $=0.3$; (b) reduced model, mode $=0.5$, mean $=0.5$; (c) full model, mode $=0.02$, mean $=0.05$; and (d) full model, mode $=0.05$, mean $=0.1$.

should expect that the power of detecting interactions will be limited given that the sample size was only one-tenth of the number of regressors. We set $c\equiv 50$ and $\tau\equiv 0.02$ with the mode and mean of the Beta hyperprior at $(0.02, 0.05)$ and $(0.05, 0.1)$. The marginal posterior probabilities of $\{\gamma_i\}$ are plotted in Figure 2(c) and (d). No regressors were selected using controlling values $\alpha_{\mathrm{L}}=0.2$ and $\alpha_{\mathrm{H}}=0.2$. According to the 3-fold rule-of-thumb, none of the interaction terms reached posterior to prior probability ratio of 3. Therefore, we concluded that no significant two-way interactions were given in the current study group.

The sample size $n=66$ was too small to explore higher-order interactions for all the 36 genes. Therefore, we examined only the six genes and all their two-way and three-way interactions, and no significant interactions were found (see the plots in Figure 3).

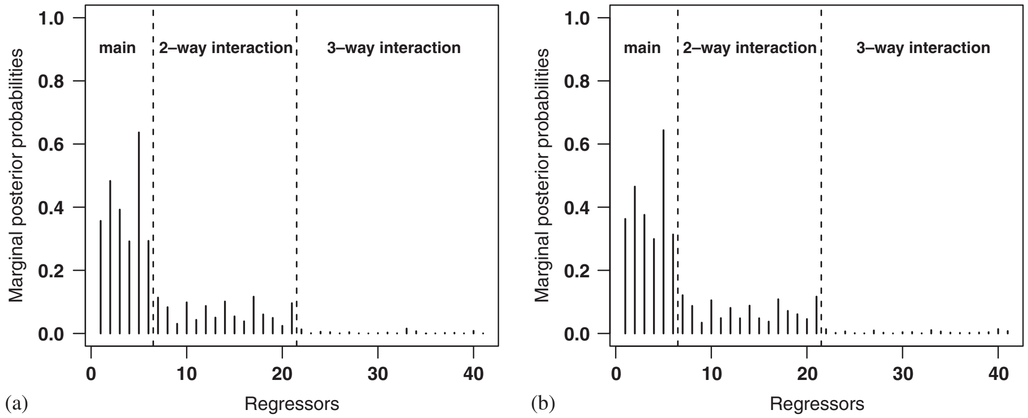

Figure 3. Marginal posterior probabilities in diffuse large-B-cell lymphoma data with higher-order interactions. The first six regressors are the main effects, the next 15 are the two-way interactions, and the rest are the three-way interactions: (a) Beta hyperprior: mean = 0.5, mode = 0.5 and (b) uniform hyperprior: U(0,1).

When Lossos *et al.* [27] performed a multiple Cox regression analysis, none of the six selected genes independently predicted overall survival at a statistically significant level. Our results agree with that of Lossos *et al.* The top six genes based on the marginal posterior probabilities from our reduced model were LMO2, BCL6, SLAM, CD38, SCYA3, and p53. Three of them (LMO2, BCL6, SCYA3) overlapped with the top six genes from Lossos *et al.*

6. CONCLUSION

The lack of statistically sound decision rules for automatic variable selection with interactions has discouraged activities in exploring important interaction effects in practice. We have proposed a general loss function framework for approaching the problem in the multiple regression setting. In particular, we have constructed loss functions that formally model the relationships between main effects and interactions while controlling the FDR. The automatic algorithm is straightforward and easily implemented using MCMC.

Different loss functions are recommended for different contexts (data and objectives). At the design stage, we cannot afford to consider a large number of interactions. Focusing on a limited number of possible interactions generates greater statistical power to test variables of primary interest. The loss function L_{LH}, which restrains higher-order terms, seems appropriate in this situation. However, most model-building procedures are carried out on observational studies, which aim to search for statistically significant associations and generate hypotheses for future studies. In these cases, exploring interactions in the full model space using the loss function L_{NC} or L_{HL} is desirable. However, we do not encourage using the loss function L_{NC}, unless enough data are available to support the existence of a special design point which, in this case, is the zero coefficient of main effect with non-zero coefficient for interaction.

Based on our simulation studies, the loss function L_{HL} (a belief that when the interactions are selected their corresponding main effects have to be forced into the model) resulted in the highest power to detect true predictors in all settings. This result was particularly obvious when the sample size is smaller than the number of regressors. In the real data analysis of the colorectal cancer study, no significant main effect was found under the additive model. Nevertheless, several interactions were found significant under the full model (main effects plus two-way interactions).

As pointed out by one referee, there are several ways to utilize the posterior distribution outputs which we obtained from the MCMC sampling. One approach would be to use the posterior distribution of total FDR and FNR at a given level of $(\alpha_{\mathrm{L}}, \alpha_{\mathrm{H}})$.

An alternative method is to use the upper quantile of posterior distribution of FDR. Given any vector of decision d, one can find whether the upper quantile of the posterior total FDR is less than the significance level. Thus, among those decisions that satisfy the significance level, the decision with the smallest posterior total FNR is optimal. However, the search will be intensive since there are many possible outcomes of d.

The optimal decision rule that we have found in this paper is a single-step procedure. However, given the marginal posterior probabilities from the MCMC output, one could also apply step-up and step-down procedures to the marginal posterior probabilities. Such a topic is beyond the scope of this paper.

As a consequence of Bayesian inference, the proposed loss functions require specifications of priors. The posterior expected loss was less sensitive to the prior when the sample size was relatively large compared with the number of regressors. When regressors are abundant, a Beta hyperprior with small mean (mean $<n/p$) and mode is appropriate. Since the mean is a prior belief of probability of being a true predictor, assuming the number of true predictors larger than the sample size (mean $>n/p$) might not make sense in the context of statistical inference.

ACKNOWLEDGEMENTS

The authors thank the editor, the associate editor, and the referees for their comments that considerably improved this article. Part of the work was supported by the National Institutes of Health through Karmanos Cancer Institute Support Grant (5 P30 CA022453-25). The work of Dr Ghosh was supported in part by the National Institutes of Health through the University of Michigan's Cancer Center Support Grant (5 P30 CA46592).

REFERENCES

1. Risch NJ. Searching for genetic determinants in the new millennium. *Nature* 2000; **405**:847–856.
2. Culverhouse R, Suarez Bk, Lin J, Reich T. A perspective on epistasis: limits of models displaying no main effect. *American Journal of Human Genetics* 2002; **70**:461–471.
3. Devlin B, Roeder k, Wasserman L. Analysis of multilocus models of association. *Genetic Epidemiology* 2003; **25**:36–47.
4. Oh C, Ye KQ, He Q, Mendell NR. Locating disease genes using Bayesian variable selection with the Haseman–Elston method. *BMC Genetics* 2003; **4**(Suppl. 1):S69.
5. Lunetta KL, Hayward LB, Segal J, Eerdewegh PV. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics* 2004; **5**:32.
6. Salanti G, Higgins JPT, White IR. Bayesian synthesis of epidemiological evidence with different combinations of exposure groups: application to a gene–gene–environment interaction. *Statistics in Medicine* 2006; **25**:4147–4163.
7. Griepentrog GL, Ryan JM, Smith LD. Linear transformations of polynomial regression models. *The American Statistician* 1982; **36**:171–174.
8. McCullagh P, Nelder JA. *Generalized Linear Models* (2nd edn). Chapman & Hall: London, 1989.
9. Peixoto JL. A property of well-formulated polynomial regression models. *The American Statistician* 1990; **44**:26–30.

10. Nelder JA. The selection of terms in response-surface models—how strong is the weak heredity principle? *The American Statistician* 1998; **52**:315–318.
11. Chipman H. Bayesian variable selection with related predictors. *The Canadian Journal of Statistics* 1996; **24**:17–36.
12. Chen W, Ghosh D, Raghunathan T, Kardia SL. A Bayesian method for finding interactions in genomic studies. *Technical Report 48*, Department of Biostatistics, University of Michigan, 2004.
13. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *Series B* 1995; **57**:289–300.
14. Ghosh D, Chen W, Raghunathan TE. The false discovery rate: a variable selection perspective. *Journal of Statistical Planning and Inference* 2006; **136**:2668–2684.
15. Genovese C, Wasserman L. Bayesian and frequentist multiple testing. In *Bayesian Statistics* Bernardo JM, Bayarri MJ, Berger JO, Dawid A, Heckerman D, Smith AFM, West M (eds). vol. 7. Oxford University Press: Oxford, 2003; 145–161.
16. Müller P, Parmigiani G, Robert C, Rousseau J. Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association* 2004; **99**:990–1001.
17. Keeney RL, Raiffa H. *Decisions with Multiple Objectives*: *Preferences and Value Tradeoffs*. Wiley: New York, 1976.
18. George EI, McCulloch RE. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 1993; **88**:881–889.
19. Delaunoit T, Alberts SR, Sargent DJ, Green E, Goldberg RM, Krook J *et al*. Chemotherapy permits resection of metastatic colorectal cancer: experience from intergroup N9741. *Annals of Oncology* 2005; **16**:425–429.
20. Goldberg RM, Sargent DJ, Morton RF, Fuchs CS, Ramanathan RK, Williamson SK *et al*. A randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer. *Journal of Clinical Oncology* 2004; **22**:23–30.
21. Chib S. Bayes inference in the tobit censored regression model. *Journal of Econometrics* 1992; **51**:79–99.
22. Wei GCG, Tanner MA. Posterior computations for censored regression data. *Journal of the American Statistical Association* 1990; **85**:829–839.
23. Dellaportas P, Smith AFM. Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistics* 1993; **42**:443–459.
24. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A *et al*. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000; **403**:503–511.
25. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC *et al*. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine* 2002; **8**:68–74.
26. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI *et al*. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine* 2002; **346**:1937–1947.
27. Lossos IS, Czerwinski DK, Alizadeh AA, Wechser MA, Tibshirani R, Botstein D, Levy R. Prediction of survival in diffuse large-B-Cell lymphoma based on the expression of six genes. *New England Journal of Medicine* 2004; **350**:1828–1837.
28. Mihara M, Erster S, Zaika A, Petrenko O, Chittenden T, Pancoska P, Moll UM. p53 has a direct apoptogenic role at the mitochondria. *Molecular Cell* 2003; **11**:577–590.
29. Gupta S, Jiang M, Anthony A, Pernis AB. Lineage-specific modulation of interleukin 4 signaling by interferon regulatory factor 4. *Journal of Experimental Medicine* 1999; **190**:1837–1848.
30. Gupta S, Anthony A, Pernis AB. Stage-specific modulation of IFN-regulatory factor 4 function by kruppel-type zinc finger proteins. *Journal of Immunology* 2001; **166**:6104–6111.
31. Polo JM, Dell’Oso T, Ranuncolo SM, Cerchietti L, Beck D, Da Silva GF *et al*. Specific peptide interference reveals BCL6 transcriptional and oncogenic mechanisms in B-cell lymphoma cells. *Nature Medicine* 2004; **10**:1329–1335.

DOI: 10.1002/sim