Penalized Logistic Regression for Detecting Gene Interactions

Mee Young Park * Trevor Hastie †

February 3, 2007

Abstract

We propose using a variant of logistic regression with L_2 regularization to fit gene-gene and gene-environment interaction models. Studies have shown that many common diseases are influenced by interaction of certain genes. Logistic regression models with quadratic penalization not only correctly characterizes the influential genes along with their interaction structures but also yields additional benefits in handling high-dimensional, discrete factors with a binary response. We illustrate the advantages of using an L_2 regularization scheme, and compare its performance with that of *Multifactor Dimensionality Reduction* and *FlexTree,* two recent tools for identifying gene-gene interactions. Through simulated and real datasets, we demonstrate that our method outperforms other methods in identification of the interaction structures as well as prediction accuracy. In addition, we validate the significance of the factors selected through bootstrap analyses.

1 Introduction

Because many common diseases are known to be affected by certain genotype combinations, there is a growing demand for methods to identify the influential genes along with their interaction structures. We propose a forward stepwise method based on penalized logistic regression. Our method primarily targets data consisting of single-nucleotide polymorphisms (SNP) measurements and a binary response variable separating the affected subjects from the unaffected ones.

Logistic regression is a standard tool for modeling effects and interactions with binary response data. However, for the SNP data here, logistic regression models have significant drawbacks:

*Ph.D. candidate, Department of Statistics, Stanford University, CA 94305. mypark@stat.stanford.edu, *tel* 16507042581

†Professor, Department of Statistics and Department of Health Research & Policy, Stanford University, CA 94305. hastie@stanford.edu

- The three-level genotype factors and their interactions can create many parameters, and with relatively small datasets, problems with overfitting arise.
- With many candidate loci, factors can be correlated leading to further degradation of the model.
- Often cells that define an interaction can be empty or nearly empty, which would require special parametrization.
- These problems are exacerbated as the interaction order is increased.

For these and other reasons, researchers have looked for alternative methods for identifying interactions.

In this paper we show that some simple modifications of standard logistic regression overcome the problems. We modify the logistic regression criterion by combining it with a penalization of the L_2 norm of the coefficients; this adjustment yields significant benefits. Because of the quadratic penalization, collinearity among the variables does not degrade fitting much, and the number of factors in the model is essentially not limited by sample size. When the levels of discrete factors are sparse or high-order interaction terms are considered, the contingency tables for the factors may easily include cells with zeros or near-zeros. Again, with the help of quadratic penalization, these situations do not diminish the stability of the fits.

We compare our method to multifactor dimensionality reduction, MDR (Ritchie, Hahn, Roodi, Bailey, Dupont, Parl & Moore 2001), a widely used tool for detecting gene interactions. The authors of MDR propose it as an alternative to logistic regression, primarily for the reasons mentioned above. Their method screens pure interactions of various orders, using cross-validation to reduce the bias of overfitting. Once an interaction is found, the inventors propose using logistic regression to tease it apart.

In the following sections, we describe and support our approach in more detail with examples and justifications. We review MDR and several other related methods in Section 2. We explore the use of penalized logistic regression in Section 3. Our methods are illustrated with simulated and real datasets in Sections 4 and 5. We conclude with a summary and possible extensions of our studies in Section 6.

2 Related Work

2.1 Multifactor Dimensionality Reduction

Multifactor dimensionality reduction (MDR), proposed by Ritchie et al. (Ritchie et al. 2001, Ritchie et al. 2003, Hahn et al. 2003, Coffey et al. 2004), is a popular technique for detecting and characterizing gene-gene/gene-environment interactions that affect complex but common genetic diseases.

2.1.1 The MDR Algorithm

MDR finds both the optimal interaction order K and the corresponding K factors that are significant in determining the disease status. The algorithm is as follows:

1. For each K, run ten-fold cross-validation to find the optimal set of K factors (described below).
2. Compare the prediction errors (on the left out set) and the consistencies (how many times out of ten-folds the optimal set of factors was selected) for different K.
3. Select the K with the smallest estimate of prediction error and/or the largest consistency. This K is the final size of the model, and the optimal set for the chosen order forms the best multifactor model.

In Step 1 above, MDR uses cross-validation to find the optimal set of factors for each K. The following steps are repeated for each cross-validation fold:

1. Construct a contingency table among every possible set of K factors.
2. Label the cells of the table *high-risk* if the cases/control ratio is greater than 1 in the training part (9/10), and *low-risk* otherwise.
3. Compute the training error for the 9/10 data, by classifying *high-risk* as a case, *low-risk* a control.
4. For the set of K factors that yields the lowest training error, compute the prediction error using the remaining 1/10.

The set of K factors that achieves the lowest training error most frequently is named the "optimal set of size K", and the largest frequency is referred to as the consistency for size K.

A strong selling point of MDR is that it can simultaneously detect and characterize multiple genetic loci associated with diseases. It searches through any levels of interaction regardless of the significance of the main effects. It is therefore able to detect high-order interactions even when the underlying main effects are statistically insignificant. However, this "strength" is also its weakness; MDR can ONLY identify interactions, and hence will suffer severely from lack of power if the real effects are additive. For example, if there are three loci active, and their effect is additive, MDR can only see them all as a three-factor interaction. Typically the power for detecting interactions decreases with K, since the number of parameters grows exponentially with K, so this is a poor approach if the real effects are additive and lower dimensional. Of course one can post-process a three-factor interaction term and find that it is additive, but the real art here is in discovering the relevant factors involved.

MDR suffers from several technical disadvantages. First, cells in high-dimensional tables will often be empty; these cells cannot be labeled based on the cases/control ratio. Second, the binary assignment (high-risk/low-risk) is highly unstable when the proportions of cases and controls are similar.

2.1.2 Dimensionality of MDR

The authors of MDR claim (Ritchie et al. 2003) that MDR reduces a p-dimensional model to a 1-dimensional model, where p is the total number of available factors. This statement is apparently based on the binary partitioning of the samples into the high-risk and low-risk groups—a one dimensional description. This characterization is flawed at several levels, because in order to produce this reduction MDR searches in potentially very high-dimensional spaces:

1. MDR searches for the optimal interaction order K.
2. MDR searches for an optimal set of K factors, among $\binom{p}{K}$ possibilities.
3. Given K factors, MDR "searches" for the optimal binary assignment of the cells of a table into *high-risk* and *low-risk*.

All these amount to an *effective dimension* or "degrees-of-freedom" that is typically much larger than one. We demonstrate, through a simulation, a more realistic assessment of the dimensionality of MDR. Here we present the results and refer the readers to Appendix A for the details of the simulation procedure.

We simulated 500 samples with 10 factors, each having 3 levels; the responses were randomly chosen to be 0/1, and thus, none of the factors was relevant to the response. Changing the order of interaction ($K = 1, 2, 3$) and the total number of available factors (from K to 10), we computed the deviance changes for the fitted models. We estimated the degrees of freedom of the models by repeating the simulation 200 times and averaging the deviance measures.

Figure 1 captures the results. The horizontal lines mark the degrees of freedom from MDR (the lower dotted line) and logistic regression (the upper dotted line) using a fixed set of factors. These are summarized as well in Table 1. For example, an MDR model with a third-order interaction of three-level factors has an effective dimension of 17.4 — above half way between the claimed 1 and the 26 of LR.

	Number of Factors K		
Method	1	2	3
MDR	1.9 (0.13)	5.6 (0.20)	17.4 (0.37)
LR	2.1 (0.14)	8.0 (0.26)	26.8 (0.53)
LR exact	2	8	26

Table 1: *The estimated degrees-of-freedom for MDR and LR, using K=1 ,2 and 3 factors (standard errors in parentheses). LR exact refers to the asymptotic exact degrees of freedom.*

2.2 Conditional Logistic Regression

Conditional logistic regression (LR) is an essential tool for the analysis of categorical factors with binary responses. Unlike MDR, LR is able to fit additive and other lower order effects

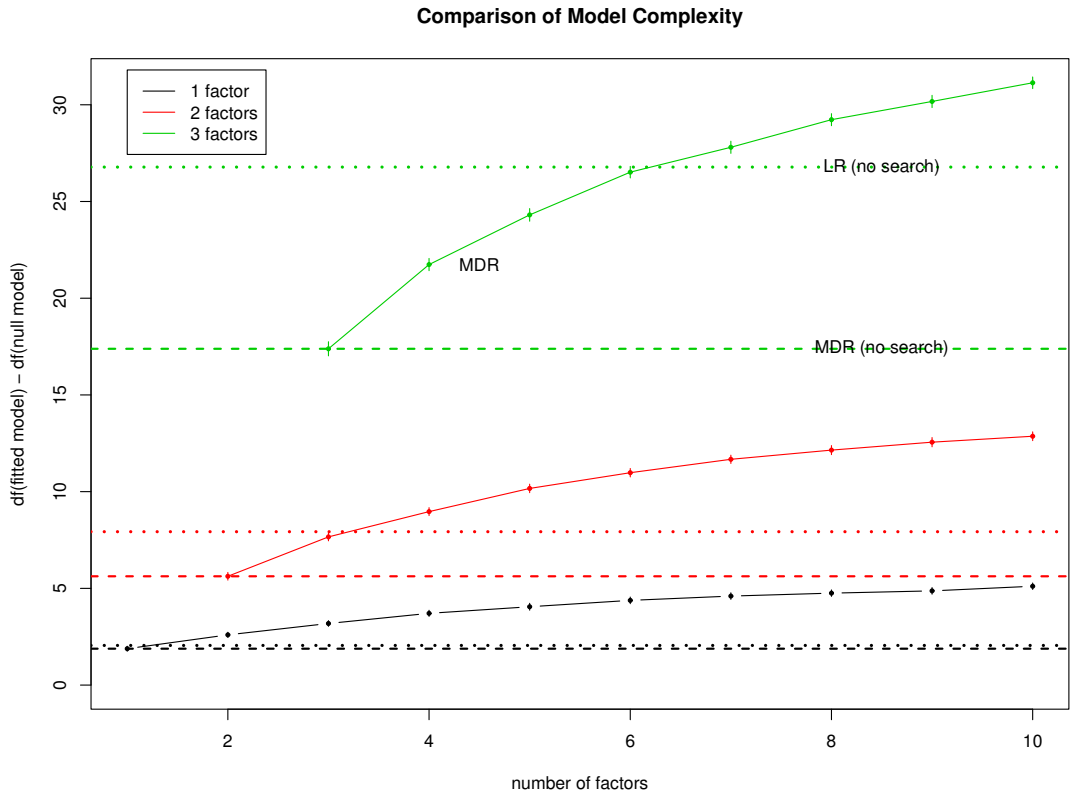

Figure 1: *Plots of the average differences in deviance between the fitted and null models: The black, red, and green solid curves represent the MDR models with the interaction orders one, two, and three, respectively. The vertical segments at the junction points are the standard error bars. As the interaction order increased, the effective degrees of freedom increased as well. In addition, each curve monotonically increased along with the number of available factors, as the optimal set of factors was searched over a larger space. The horizontal lines mark the degrees of freedom from MDR (the lower dotted line) and logistic regression (the upper dotted line) without searching (we used a fixed set of factors, so that there was no effect due to searching for an optimal set of factors).*

as well as full-blown interactions. Therefore, LR can yield a more precise interpretation that distinguishes the presence of additive effects from the presence of interaction effects. In fact, the users of MDR fit LR models using the factors selected by MDR precisely for this reason; to simplify the high-order interactions into its component effects. LR is sometimes criticized due to the difficulties of estimating a large number of parameters with a relatively small number of samples (Ritchie et al. 2001); however, we provide a solution to overcome this drawback. Biologists (Coffey et al. 2004, for example) have shown that LR performs as well as other methods in cases where it is able to be fit.

2.3 FlexTree

Huang et al. (2004) proposed a tree-structured learning method, *FlexTree,* to identify the genes related to the cause of complex diseases along with their interactions. It is a rather complex procedure that aims to build a tree with splits in the form of a linear combination of multiple factors. Beginning from the root node that contains all the observations, each node is recursively split into two daughter nodes, through the following steps:

1. Use backward shaving to select the optimal set of predictors for splitting the specific node. For the backward shaving, form a decreasing series of candidate subsets based on the bootstrapped scores. Then determine the best subset among the series that yields the largest cross-validated impurity measure.

2. Perform a permutation test to see if the linear relationship between the selected subset of predictors and the outcome is strong enough. If so, go to the next step. If not, stop splitting the node.

3. Use the selected subset of predictors to compute the regression coefficients ($\hat{\beta}$) and the splitting threshold (C) such that a binary split is determined based on $\mathbf{x}'\hat{\beta} \geq C$. The optimal scoring method is used for estimating β, and C is chosen to maximize the resulting Gini index for the node.

Huang et al. (2004) compared FlexTree to other methods such as CART, QUEST, logic regression, bagging, MART, and random forest; they showed that FlexTree performed better than or as well as these competing methods. Using a very similar dataset, we compared the performance of our method with that of FlexTree (in Section 5).

3 Penalized Logistic Regression

The generic logistic regression model has the form

$$\log \frac{Pr(Y=1|X)}{Pr(Y=0|X)} = \beta_0 + X^T\beta, \tag{1}$$

where X is a vector of predictors (typically dummy variables derived from factors, in the present setting). Logistic regression coefficients are typically estimated by maximum-likelihood (McCullagh & Nelder 1989); in fact the deviance (16) that we used in Section 2.1.2 is twice the negative log-likelihood. Here we maximize the log-likelihood subject to a size constraint on L_2 norm of the coefficients (excluding the intercept) as proposed in Lee & Silvapulle (1988) and Le Cessie & Van Houwelingen (1992). This amounts to minimizing the following equation:

$$L(\beta_0, \beta, \lambda) = -l(\beta_0, \beta) + \frac{\lambda}{2}\|\beta\|_2^2, \tag{2}$$

where l indicates the binomial log-likelihood, and λ is a positive constant. The coefficients are regularized in the same manner as in ridge regression (Hoerl & Kennard 1970). The importance of the quadratic penalty, particularly in our application, will be elaborated in subsequent sections.

To fit penalized logistic regression models, we repeat the Newton-Raphson steps, which result in the *iteratively reweighted ridge regressions* (IRRR) algorithm:

$$\begin{aligned}
\boldsymbol{\beta}^{new} &= \boldsymbol{\beta}^{old} - (\frac{\delta^2 L}{\delta\boldsymbol{\beta}\delta\boldsymbol{\beta}^T})^{-1}\frac{\delta L}{\delta\boldsymbol{\beta}} && (3)\\
&= (\mathbf{X}^T\mathbf{W}\mathbf{X} + \boldsymbol{\Lambda})^{-1}\mathbf{X}^T\mathbf{W}\{\mathbf{X}\boldsymbol{\beta}^{old} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})\} && (4)\\
&= (\mathbf{X}^T\mathbf{W}\mathbf{X} + \boldsymbol{\Lambda})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}. && (5)
\end{aligned}$$

$\mathbf{X}$ is the $n \times (p+1)$ matrix of the predictors (n and p are the numbers of the samples and the predictors, respectively); $\mathbf{y}$ is the vector of 0/1 responses; $\mathbf{p}$ is the vector of probability estimates that the responses are equal to 1; $\mathbf{W}$ is the diagonal matrix with the diagonal elements $p_i(1-p_i)$ for $i = 1, \ldots, n$; $\boldsymbol{\Lambda}$ is the diagonal matrix with the diagonal elements $\{0, \lambda, \ldots, \lambda\}$; and $\mathbf{z} = \mathbf{X}\boldsymbol{\beta}^{old} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})$ is the current *working* response in the IRRR algorithm.

As a result of the quadratic penalization, the norm of the coefficient estimates is smaller than in the case of regular logistic regression; however, none of the coefficients is zero. As in ridge regression, the amount of shrinkage that gets applied to each coefficient depends on the variance of the corresponding factor. This analogy to ridge regression is easily seen from (3)-(5).

Using the values from the final Newton-Raphson step of the IRRR algorithm, we estimate the effective degrees of freedom of the model (Hastie & Tibshirani 1990) and the variance of the coefficient estimates (Gray 1992). The effective degrees of freedom are approximated by

$$df(\lambda) = tr[(\mathbf{X}^T\mathbf{W}\mathbf{X} + \boldsymbol{\Lambda})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{X}], \tag{6}$$

where $\mathbf{W}$ is obtained from the final step of the algorithm. This representation is based on similar ideas to those described in Appendix A. The variance of the coefficients is also

estimated from the final iteration:

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= Var[(\mathbf{X}^T\mathbf{W}\mathbf{X}+\boldsymbol{\Lambda})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}] && (7)\\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X}+\boldsymbol{\Lambda})^{-1}Var[\mathbf{X}^T(\mathbf{y}-\mathbf{p})](\mathbf{X}^T\mathbf{W}\mathbf{X}+\boldsymbol{\Lambda})^{-1} && (8)\\ &= (\frac{\delta^2 L}{\delta\boldsymbol{\beta}\delta\boldsymbol{\beta}^T})^{-1}I(\boldsymbol{\beta})(\frac{\delta^2 L}{\delta\boldsymbol{\beta}\delta\boldsymbol{\beta}^T})^{-1}, && (9)\end{aligned}$$

where $I(\boldsymbol{\beta})$ denotes the information in $\mathbf{y}$. This is referred to as a *sandwich estimate* (Gray 1992).

We now elaborate on the use of penalized logistic regression specifically as it relates to our problem.

3.1 Advantages of Quadratic Penalization

Using quadratic regularization with logistic regression has a number of attractive properties.

1. When we fit interactions between categorical factors, the number of parameters can grow large. The penalization nevertheless enables us to fit the coefficients in a stable fashion.

2. We can code factors in a symmetric fashion using dummy variables, without the usual concern for multicollinearity. (In Section 3.5, we introduce a missing value imputation method taking advantage of this coding scheme.)

3. Zero cells are common in multi-factor contingency tables. These situations are handled gracefully.

Since quadratic regularization overcomes collinearity amongst the variables, a penalized logistic regression model can be fit with a large number of factors or high-order interaction terms. The sample size does not limit the number of parameters. In Section 3.2, we illustrate our variable selection strategy; a growing number of variables in the model is not detrimental to the variable search.

The quadratic penalty makes it possible to code each level of a factor by a dummy variable, yielding coefficients with direct interpretations. Each coefficient reveals the significance of a particular level of a factor. This coding method cannot be applied to regular logistic regression because the dummy variables representing a factor are perfectly collinear (they sum to one). To overcome this, one of the levels is omitted, or else the levels of the factors are represented as contrasts.

It turns out that the penalized criterion (2) creates the implicit constraint that the coefficients of the dummy variables representing any discrete factor/interaction of factors must sum to zero. Consider the model

$$\log\frac{Pr(Y=1|D)}{Pr(Y=0|D)} = \beta_0 + D^T\beta, \qquad (10)$$

where D is a vector of dummy variables that represent the levels of a three-level categorical factor. As can be seen from (10), adding a constant vector to β and subtracting the same constant from β_0 would not change the probability estimate. However, because our criterion minimizes $\|\beta\|_2$, the coefficients are identifiable in such a way that the elements of β sum to zero. Given a dataset of n observations (d_i, y_i), we differentiate the objective function (2) with respect to the coefficients and obtain:

$$\frac{\delta L}{\delta \beta_0} = 0 \iff \sum_{i=1}^{n}(y_i - p_i) = 0, \tag{11}$$

$$\frac{\delta L}{\delta \beta} = 0 \iff \sum_{i=1}^{n}(y_i - p_i)d_i = \lambda\beta. \tag{12}$$

These equations, in turn, imply $\sum_{j=1}^{3}\beta_j = 0$. Zhu & Hastie (2004) explored this property of the L_2 penalization in (multinomial) penalized logistic regression using continuous factors. We can learn more from (12):

- Higher order interactions have particular marginal constraints. For example, consider an interaction between factor 1 and 2:

$$\beta_j^1 = \sum_k \beta_{jk}^{12} \; \forall j; \tag{13}$$

$$\beta_k^2 = \sum_j \beta_{jk}^{12} \; \forall k. \tag{14}$$

  We can see this by summing over appropriate subsets of the equations in (12).

- Each of these imply that $\sum_{j,k}\beta_{jk}^{12} = 0$.

- A generalization is that for any penalized interaction term, if the coefficients of a marginal is *not* penalized, the sums as in (13) are zero.

When a column of $\mathbf{X}$ is so unbalanced that it contains no observations at a particular level (or combination of levels), the corresponding dummy variable is zero for all n observations. This phenomenon is common in SNP data because one allele of a locus can easily be prevalent over the other allele on the same locus. The lack of observations for certain levels of factors occurs even more frequently in high-order interaction models. We cannot fit a regular logistic regression model with an input matrix that contains a column of zeros. However, when logistic regression is accompanied by any small amount of quadratic penalization, the coefficient of the zero column will automatically be zero.

We demonstrate this for a simple two-way interaction term in a model. As in (12),

$$\frac{\delta L}{\delta \beta_{jk}^{12}} = 0 \Longleftrightarrow \lambda\beta_{jk}^{12} = m_{jk} - \frac{1}{1 + e^{-(\beta_0 + \beta_j^1 + \beta_k^2 + \beta_{jk}^{12})}} n_{jk}, \tag{15}$$

where n_{jk} is the number of observations with $X_1 = j$ and $X_2 = k$, m_{jk} is the number among these with $Y = 1$, and β_j^1 is the coefficient for the jth level of variable 1, etc. The equivalence implies that if $n_{jk} = 0$, then $m_{jk} = 0$, and, thus, $\hat{\beta}_{jk}^{12} = 0$ for any $\lambda > 0$. An analogous equality holds at any interaction order.

3.2 Variable Selection: Forward Stepwise Procedure

Penalizing the norm of the coefficients results in a smoothing effect in most cases. However, with an L_2 penalization, none of the coefficients is set to zero unless the distribution of the factors is extremely sparse as illustrated in the previous section. For prediction accuracy and interpretability, we often prefer using only a subset of the features in the model, and thus we need to include a method for variable selection. For that purpose we use a classic approach: forward selection, followed by backward deletion.

In each forward step, a factor/interaction of factors is added to the model. A preset total number of forward steps are repeated. In the subsequent backward steps, a factor/interaction of factors is deleted, beginning with the final (and largest) model from the forward steps. Backward deletion continues until only one factor remains in the active set (the set of variables in the model). The choice of factor to be added or deleted in each step is based on the cost-complexity statistic $C = \textit{deviance} + \textit{cp} \times \textit{df}$, where *cp* is *complexity parameter*. Popular choices are $\textit{cp} = 2$ and $\textit{cp} = \log(\textit{sample size})$ for AIC and BIC, respectively.

When adding or deleting variables, we obey the hierarchy principle: when an interaction of multiple factors is in the model, the lower order factors comprising the interaction should also be present. This is symmetric in the two lower-order factors; we in fact implement a less stringent asymmetric version. To allow interaction terms to enter the model more easily, we modify the convention, such that any factor/interaction of factors in the active set can form a new interaction with any other single factor, even when the single factor is not yet in the active set. This asymmetric hierarchy was proposed by Friedman (1991) in the context of MARS models. To add more flexibility, an option we allow is to provide all possible second-order interactions as well as main-effect terms as candidate factors at the beginning of the forward steps.

In the backward deletion process, we again respect the asymmetric hierarchy; not all components (of lower-order) of an interaction term can be dropped before the interaction term. As the size of the active set is reduced monotonically, the backward deletion process produces a series of models, from the most complex model to the NULL model. Using the list of corresponding scores, we select the model size that generated the minimum score C.

3.3 Variable Selection: Forward Stagewise Procedure

Our quadratic penalization scheme has an unwelcome side effect, particularly when the regularization parameter λ is large. Suppose factor X_1 has a strong effect, and factor X_2 has no effect, either as a main effect or interaction with X_1. With only X_1 in the model, and a sizeable quadratic penalty parameter λ, its coefficients β_j^1 are not allowed to enter

at full strength. If we include the interaction term $X_1 \times X_2$, its coefficients obey (13) $\sum_k \beta_{jk}^{12} = \beta_j^1 \ \forall j$, and hence they can augment the reduced strength main effects. Since, for example, $\delta^2 > (\delta/2)^2 + (\delta/2)^2$, it is easy to see that quadratic penalization *favors* breaking up a large main effect coefficient into a sum of smaller interaction pieces. This suggests that interaction terms might be included in the model to compensate for the regularized main effects, rather than for the purpose intended.[1]

We often use a small value of λ, largely as a device for controlling numerical stability; in these cases this effect will be negligible. When larger values are used, we propose an alternative and more restrictive forward *stagewise* selection approach.

That is, when a variable is added to the model, the coefficients for all the previously-included factors are fixed, and only the coefficients for the new variable are penalized. This prevents the earlier coefficients from re-adjusting themselves to take advantage of the phenomenon described above.

We compare the forward stagewise and stepwise procedures in Section 5; the sets of factors selected by the two procedures are very similar and often identical, because the factors that enter the model in the forward stagewise procedure are often in the same sequence as in the forward selection steps. Our software offers both choices, but based on our experience so far the original stepwise procedure is our preferred choice.

3.4 Choosing the Regularization Parameter λ

Here we explore the smoothing effect of L_2 penalization, mentioned briefly in Section 3.2. When building factorial models with interactions, there is always a risk of overfitting the data, even with a selected subset of the features. In addition to the advantages of using quadratic regularization emphasized in Section 3.1, it can also be used to smooth a model and thus control its effective size or degrees of freedom (6). As heavier regularization is imposed by increasing λ, the deviance of the fit increases (the fit degrades), but the variance (7)-(9) of the coefficients and the effective degrees of freedom (6) of the model decrease. As a result, when the model size is determined by AIC/BIC, a larger value of λ tends to choose a model with more variables and allow complex interaction terms to join the model more easily.

We used simulation to illustrate the patterns of models selected by varying λ, and for guidance in choosing an appropriate value. We ran three sets of simulation analyses, for each one generating data with a different magnitude of interaction effect. For all three cases, we generated datasets consisting of a binary response and six categorical factors with three levels. Only the first two of the six predictors affected the response, with the conditional probabilities of belonging to class 1 as in the tables below. (AA,Aa,aa) and (BB,Bb,bb) are the possible genotypes for the first two factors; $P(A) = P(B) = 0.5$, meaning that the proportions of the alleles A and B in population are assumed to be the same as the alleles a and b, respectively. Figure 2 displays the log-odds for class 1, for all possible combinations

[1] We thank one of the referees for alerting us to this phenomenon.

of levels of the first two factors; the log-odds are additive for the first model, while the next two show interaction effects.

Additive Model

$P(A) = P(B) = 0.5$

	BB	Bb	bb
AA	0.845	0.206	0.206
Aa	0.206	0.012	0.012
aa	0.206	0.012	0.012

Interaction Model I

$P(A) = P(B) = 0.5$

	BB	Bb	bb
AA	0.145	0.206	0.206
Aa	0.206	0.012	0.012
aa	0.206	0.012	0.012

Interaction Model II

$P(A) = P(B) = 0.5$

	BB	Bb	bb
AA	0.045	0.206	0.206
Aa	0.206	0.012	0.012
aa	0.206	0.012	0.012

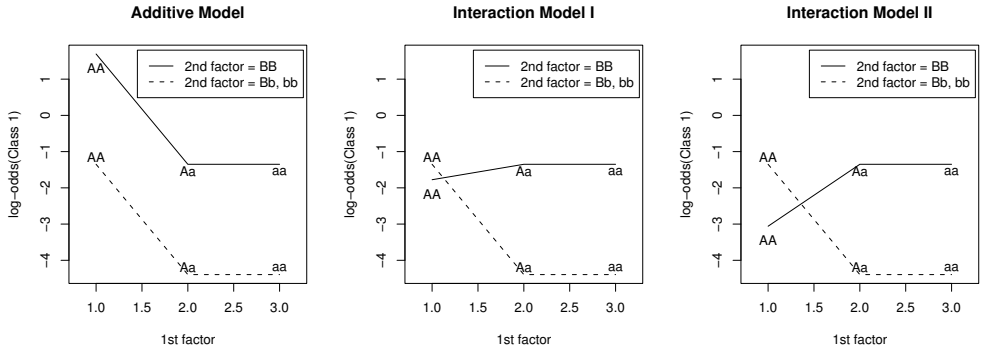

Figure 2: *The patterns of log-odds for class 1, for different levels of the first two factors*

For each model, we generated 30 datasets of size 100, with balanced class labels. Then, we applied our procedure with $\lambda = \{0.01, 0.5, 1, 2\}$, for each λ selecting a model based on BIC. Table 2 summarizes how many times $\mathbf{A} + \mathbf{B}$ (the additive model with the first two factors) and $\mathbf{A} * \mathbf{B}$ (the interaction between the first two factors) were selected. The models that were not counted in the table include $\mathbf{A}$, $\mathbf{B}$, or the ones with the terms other than $\mathbf{A}$ and $\mathbf{B}$. Given that $\mathbf{A} + \mathbf{B}$ is the true model for the first set while $\mathbf{A} * \mathbf{B}$ is appropriate for the second and the third, ideally we should be fitting with small values of λ for the additive model but increasing λ as stronger interaction effects are added.

λ		0.01	0.5	1	2
Additive Model	$\mathbf{A} + \mathbf{B}$	28/30	26/30	26/30	22/30
	$\mathbf{A} * \mathbf{B}$	0/30	0/30	0/30	5/30
Interaction Model I	$\mathbf{A} + \mathbf{B}$	16/30	14/30	10/30	6/30
	$\mathbf{A} * \mathbf{B}$	11/30	14/30	18/30	23/30
Interaction Model II	$\mathbf{A} + \mathbf{B}$	6/30	5/30	3/30	1/30
	$\mathbf{A} * \mathbf{B}$	20/30	24/30	26/30	27/30

Table 2: *The number of times that the additive and the interaction models were selected.* $\mathbf{A} + \mathbf{B}$ *is the true model for the first set while* $\mathbf{A} * \mathbf{B}$ *is the true model for the second and the third.*

We can cross-validate to choose the value of λ; for each fold, we obtain a series of optimal models (based on AIC/BIC) corresponding to the candidate values of λ and compute the log-likelihoods using the omitted fold. Then we choose the value of λ that yields the largest average (cross-validated) log-likelihood. We demonstrate this selection strategy in Section 4.

3.5 Missing Value Imputation

The coding method we implemented suggests an easy, but reasonable, method of imputing any missing values. If there are any samples lacking an observation for a factor X_j, we then compute the sample proportions of the levels of X_j among the remaining samples. These proportions, which are the expected values of the dummy variables, are assigned to the samples with missing cells.

In this scheme, the fact that the dummy variables representing any factor sum to 1 is retained. In addition, our approach offers a smoother imputation than does filling the missing observations with the level that occurred most frequently in the remaining data. Through simulations in Section 4, we show that this imputation method yields a reasonable result.

4 Simulation Study

To compare the performance of penalized logistic regression to that of MDR under various settings, we generated three epistatic models and a heterogeneity model, some of which are based on the suggestions in Neuman & Rice (1992). Each training dataset contained 400 samples (200 cases and 200 controls) and 10 factors, only two of which were significant. Three levels of the two significant factors were distributed so that the conditional probabilities of being diseased were as in the tables below; the levels of the remaining eight insignificant factors were in Hardy-Weinberg equilibrium. For all four examples, the overall proportion of the diseased population was 10%.

Epistatic Model I

$P(A) = 0.394,\ P(B) = 0.340$

	BB	Bb	bb
AA	0.7	0.7	0
Aa	0.7	0	0
aa	0	0	0

Epistatic Model II

$P(A) = 0.450,\ P(B) = 0.283$

	BB	Bb	bb
AA	0	0.4	0.4
Aa	0.4	0	0
aa	0.4	0	0

Epistatic Model III

$P(A) = 0.3,\ P(B) = 0.3$

	BB	Bb	bb
AA	0.988	0.5	0.5
Aa	0.5	0.01	0.01
aa	0.5	0.01	0.01

Heterogeneity Model

$P(A) = 0.512,\ P(B) = 0.303$

	BB	Bb	bb
AA	0.415	0.35	0.35
Aa	0.1	0	0
aa	0.1	0	0

Figure 3 contains the plots of the log-odds for all the conditional probabilities in the tables. (Zeros are replaced by 0.001 to compute the log-odds.) As can be seen, we designed the third epistatic model so that the log-odds are additive (the odds are multiplicative) in the first two factors; the interaction effect is more obvious in the first two epistatic models than in the heterogeneity model. Our distinction of the heterogeneity and epistatic models is based on Vieland & Huang (2003) and Neuman & Rice (1992). We discuss how it is different from the additive/interaction scheme in logistic regression in Appendix B.

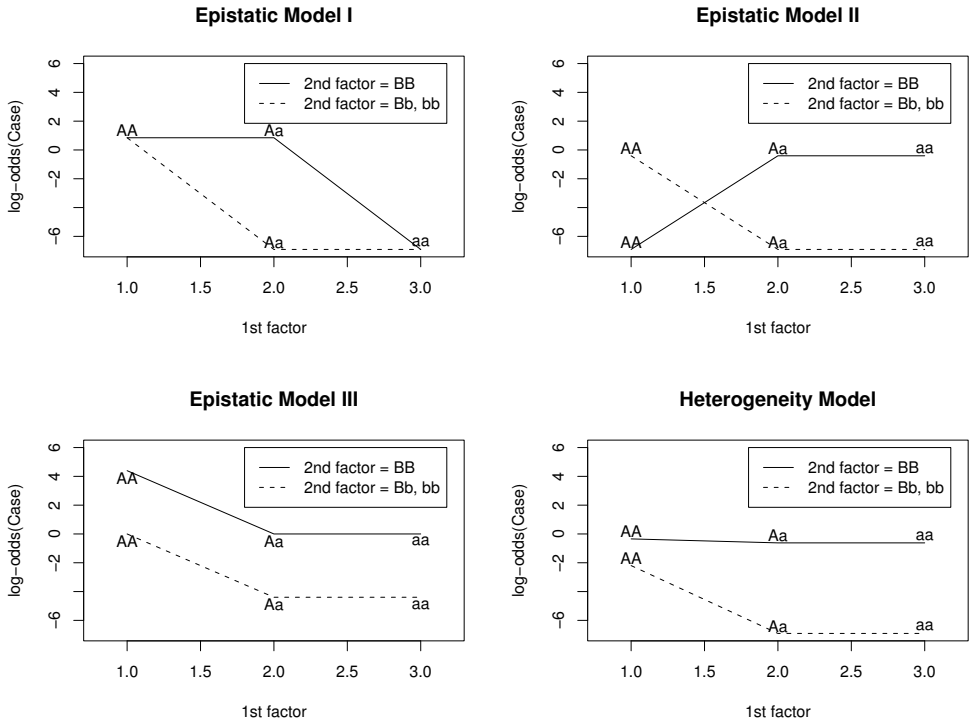

Figure 3: *The patterns of the log-odds for case, for different levels of the first two factors*

In addition to this initial simulation, we added noise to the data as described in Ritchie et al. (2003). The data were perturbed to create the following errors:

1. Missing cells (MS): For 10% of the samples, one of the significant factors is missing.

2. Genetic heterogeneity (GH): For 50% of the cases, the third and the fourth factors, instead of the first two, are significant.

We used the initial data with no error and the perturbed data to compare the prediction accuracy and power in detecting the significant factors between our method and MDR.

Under each scenario, we simulated thirty sets of training and test datasets. For each training set, we selected the regularization parameter λ through cross-validation, and using the chosen λ, built a model based on the BIC criterion. For each cross-validation, we provided

candidate values of λ in an adaptive way. We first applied a small value, $\lambda = 10^{-5}$, to the whole training dataset and achieved models of different sizes from the backward deletion. Based on the series of models, we defined a set of reasonable values for the effective degrees of freedom. Then we computed the values of λ that would reduce the effective degrees of freedom of the largest model to the smaller values in the set.

We measured the prediction errors by averaging the thirty test errors. Table 3 summarizes the prediction accuracy comparison of penalized logistic regression and MDR; the standard errors of the error estimates are parenthesized. The table shows that for both methods, the error rates increase when the data contain errors. The prediction accuracies are similar between the two methods, although MDR yields slightly larger error rates in most situations.

Model		**No error**	**MS**	**GH**
Epistatic I	PLR	0.023(0.001)	0.025(0.001)	0.111(0.002)
	MDR	0.023(0.001)	0.029(0.001)	0.131(0.002)
Epistatic II	PLR	0.085(0.001)	0.092(0.001)	0.234(0.004)
	MDR	0.084(0.001)	0.093(0.002)	0.241(0.004)
Epistatic III	PLR	0.096(0.002)	0.099(0.002)	0.168(0.003)
	MDR	0.097(0.002)	0.105(0.002)	0.192(0.005)
Heterogeneity	PLR	0.144(0.002)	0.146(0.002)	0.304(0.004)
	MDR	0.148(0.002)	0.149(0.002)	0.310(0.004)

Table 3: *The prediction accuracy comparison of PLR and MDR (the standard errors are parenthesized): The prediction accuracies are similar between the two methods, although MDR yields slightly larger error rates in most situations.*

Table 4 contains the numbers counting the cases (out of 30) for which the correct factors were identified. For PLR, the number of cases for which the interaction terms were also selected is parenthesized; the numbers vary reflecting the magnitude of interaction effect imposed in these four models as shown in Figure 3.

Model		**No error**	**MS**	**GH**
Epistatic I	PLR	30(27)	30(30)	30(29)
	MDR	30	29	30
Epistatic II	PLR	30(29)	30(28)	30(25)
	MDR	27	29	16
Epistatic III	PLR	30(1)	30(2)	30(2)
	MDR	27	29	26
Heterogeneity	PLR	30(10)	30(10)	30(8)
	MDR	23	27	5

Table 4: *The number of cases (out of 30) for which the correct factors were identified. For PLR, the number of cases that included the interaction terms is in the parentheses.*

For the heterogeneity model, main effects exist for both of the two significant factors. In addition, as one is stronger than the other, MDR was not successful in identifying them

simultaneously even for the data with no error, as shown in Table 4. In the case of the heterogeneity model or the second epistatic model, MDR suffered from a decrease in power, especially with GH perturbations. When GH perturbations were added to the second epistatic model, MDR correctly specified the four factors only 16 out of 30 times, while our method did so in all 3 simulations. These results show that the penalized logistic regression method is more powerful than MDR, especially when multiple sets of significant factors exist; in these situations, MDR often identifies only a subset of the significant factors.

5 Real Data Example

5.1 Hypertension Dataset

We compared our method to Flextree and MDR using the data from the SAPPHIRe (Stanford Asian Pacific Program for Hypertension and Insulin Resistance) project. The goal of the SAPPHIRe project was to detect the genes that predispose individuals to hypertension. A similar dataset was used in Huang et al. (2004) to show that the FlexTree method outperforms many competing methods. The dataset contains the menopausal status and the genotypes on 21 distinct loci of 216 hypotensive and 364 hypertensive Chinese women. The subjects' family information is also available; samples belonging to the same family are included in the same cross-validation fold for all the analyses.

5.1.1 Prediction performance

We applied five-fold cross-validation to estimate the misclassification rates using penalized logistic regression with forward stepwise variable selection, FlexTree, and MDR. For penalized logistic regression, a complexity parameter was chosen for each fold through an internal cross-validation. For MDR, we used internal cross-validations to select the most significant sets of features.

Huang et al. (2004) initially used an unequal loss for the two classes: misclassifying a hypotension sample was twice as costly as misclassifying a hypertension sample. We fit penalized logistic regression, FlexTree and MDR with an equal as well as an unequal loss. For MDR with an unequal loss, we used *1/(1+2)* as threshold when labeling the cells of the tables, instead of *1/2* as in the usual cases of equal loss.

The results are compared in Table 5. Penalized logistic regression achieved lower misclassification cost than other methods with either loss function. When an equal loss was used, FlexTree and MDR generated highly unbalanced predictions, assigning most samples to the larger class. Although penalized logistic regression also achieved a low specificity, it was not so serious as in the other two methods.

Figure 4 shows the receiver operating characteristic (ROC) curves for penalized logistic regression with an unequal (left panel) and an equal (right panel) loss function. For both plots, vertical and horizontal axes indicate the sensitivity and the specificity respectively. Because penalized logistic regression yields the predicted probabilities of a case, we could

Method (loss)	Miscost	Sensitivity	Specificity
PLR (unequal)	$141 + 2 \times 85 = 311$	$223/364 = 0.613$	$131/216 = 0.606$
FlexTree (unequal)	$129 + 2 \times 105 = 339$	$235/364 = 0.646$	$111/216 = 0.514$
MDR (unequal)	$122 + 2 \times 114 = 350$	$242/364 = 0.665$	$102/216 = 0.472$
PLR (equal)	$72 + 139 = 211$	$292/364 = 0.802$	$77/216 = 0.356$
FlexTree (equal)	$61 + 163 = 224$	$303/364 = 0.832$	$53/216 = 0.245$
MDR (equal)	$73 + 163 = 236$	$291/364 = 0.799$	$53/216 = 0.245$

Table 5: *Comparison of prediction performance among different methods.*

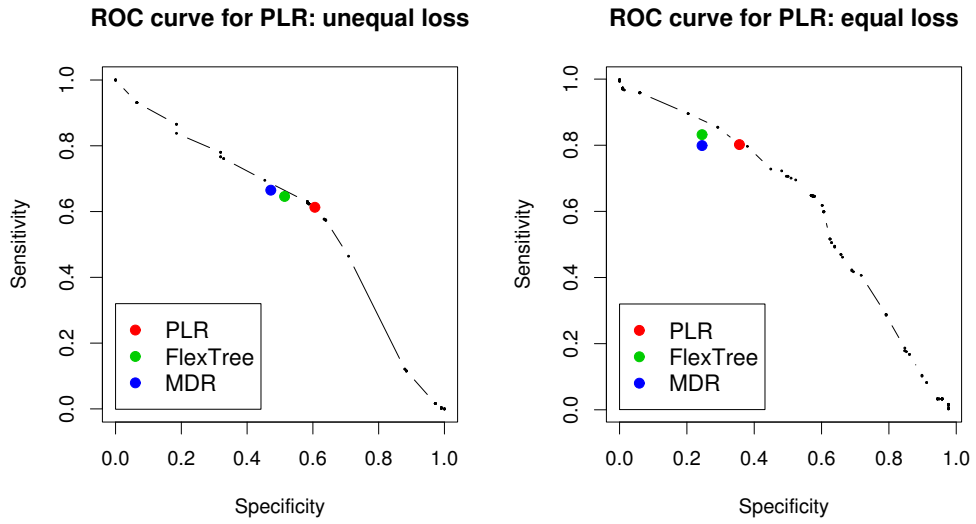

Figure 4: *Receiver operating characteristic (ROC) curves for penalized logistic regression with an unequal (left panel) and an equal (right panel) loss function: Penalized logistic regression would achieve a higher sensitivity (specificity) than other methods if the specificity (sensitivity) were fixed the same as theirs.*

compute different sets of sensitivity and specificity by changing the classification threshold between 0 and 1. The red dots on the curves represent the values we achieved with the usual threshold 0.5. The green dots corresponding to Flextree and the blue dots corresponding to MDR are all located toward the lower left corner, away from the ROC curves. In other words, penalized logistic regression would achieve a higher sensitivity (specificity) than other methods if the specificity (sensitivity) were fixed the same as theirs.

5.1.2 Bootstrap analysis of the feature selection

Applying our forward stepwise procedure to the whole dataset yields a certain set of significant features as listed in the first column of Table 6. However, if the data were perturbed, a different set of features would be selected. Through a bootstrap analysis (Efron & Tibshirani 1993), we provide a measure of how likely the features were to be selected and examine what other factors could have been preferred.

Factors selected from the whole data	Frequency
menopause	299/300
MLRI2V	73/300
Cyp11B2x1INV $\times$ *MLRI2V*	3/300
KLKQ3E	29/300
KLKQ3E $\times$ *Cyp11B2x1INV* $\times$ *MLRI2V*	10/300
AGT2R1A1166C $\times$ *menopause*	106/300

Table 6: *Significant factors selected from the whole dataset (left column) and their frequencies in 300 bootstrap runs (right column).*

We illustrate the bootstrap analysis using a fixed value of λ. For each of $B = 300$ bootstrap datasets, we ran a forward stepwise procedure with $\lambda = 0.25$, which is a value that was frequently selected in previous cross-validation. At the end of the B bootstrap runs, we counted the frequency for every factor/interaction of factors that has been included in the model at least once. The second column of Table 6 contains the counts for the corresponding features; some of them were rarely selected. Table 7 lists the factors/interactions of factors that were selected with relatively high frequencies.

Not all of the commonly selected factors listed in Table 7 were included in the model when we used the whole dataset. It is possible that some factors/interactions of factors were rarely selected simultaneously because of a strong correlation among them. To detect such instances, we propose using the co-occurrence matrix (after normalizing for the individual frequencies) among all the factors/interactions of factors listed in Table 7 as a dissimilarity matrix and applying hierarchical clustering. Then any group of factors that tends not to appear simultaneously would form tight clusters.

Using the 11 selected features in Table 7, we first constructed the 11×11 co-occurrence matrix, so that the (i, j) element was the number of the bootstrap runs in which the *i-th* and the *j-th* features were selected simultaneously. Then we normalized the matrix by dividing

Factor	Frequency	Interaction of factors	Frequency
menopause	299/300	*menopause* × *AGT2R1A1166C*	106/300
MLRI2V	73/300	*menopause* × *ADRB3W1R*	48/300
AGT2R1A1166C	35/300	*menopause* × *Cyp11B2x1INV*	34/300
HUT2SNP5	34/300	*menopause* × *Cyp11B2-5'aINV*	33/300
PTPN1i4INV	34/300	*menopause* × *AVPR2G12E*	31/300
PPARG12	30/300		

Table 7: *Factors/interactions of factors that were selected in 300 bootstrap runs of forward stepwise procedure with relatively high frequencies.*

the (i, j) entry by the number of bootstrap runs in which either the *i-th* or the *j-th* feature was selected. That is, denoting the (i, j) entry as M_{ij}, we divided it by $M_{ii} + M_{jj} - M_{ij}$ for every i and j.

As we performed hierarchical clustering with the normalized co-occurrence distance measure, *PTPN1i4INV* and *MLRI2V* were in a strong cluster: they were in the model simultaneously for only two bootstrap runs. Analogously, *menopause* × *AGT2R1A1166C* and *AGT2R1A1166C* appeared 106 and 35 times respectively, but only twice simultaneously. For both clusters, one of the elements was selected in our model (Table 6) while the other was not. Hence, the pairs were presumably used as alternatives in different models.

To compare the forward stagewise variable selection scheme to the forward stepwise method that we have used so far, we again summarize the factors/interactions of factors that were selected in 300 bootstrap runs of forward stagewise procedure in Table 8. Although the forward stagewise procedure tends to select single factors more frequently, and the two-way interaction terms less frequently compared to the previous scheme, the two methods share common factors/interactions of factors. For instance, *menopause* and *MLRI2V* were the two most frequent main effect terms; and *menopause* × *Cyp11B2x1INV* and *menopause* × *AGT2R1A1166C* were among the most common interaction terms.

Factor	Frequency	Interaction of factors	Frequency
menopause	272/300	*menopause* × *Cyp11B2x1INV*	68/300
MLRI2V	152/300	*menopause* × *PPARG12*	52/300
PTPN1x9INV	83/300	*menopause* × *AGT2R1A1166C*	38/300
KLKQ3E	62/300	*AGT2R1A1166C* × *MLRI2V*	36/300
AGT2R2C1333T	45/300		
AGT2R1A1166C	31/300		

Table 8: *Factors/interactions of factors that were selected in 300 bootstrap runs of forward stagewise procedure with relatively high frequencies.*

5.2 Bladder Cancer Dataset

We show a further comparison of different methods with another dataset, which was used by Hung et al. (2004) for a case-control study of bladder cancer. The dataset consisted of genotypes on 14 loci and the smoke status of 201 bladder cancer patients and 214 controls.

5.2.1 Prediction performance

We compared the prediction error rate of penalized logistic regression with those of Flextree and MDR through five-fold cross-validation. As summarized in Table 9, penalized logistic regression achieved higher sensitivity and specificity than Flextree, and better balanced class predictions than MDR.

Method	Misclassification error	Sensitivity	Specificity
PLR	$147/415 = 0.354$	$122/201 = 0.607$	$146/214 = 0.682$
Flextree	$176/415 = 0.424$	$107/201 = 0.532$	$132/214 = 0.617$
MDR	$161/415 = 0.388$	$137/201 = 0.682$	$117/214 = 0.547$

Table 9: *Comparison of prediction performance among different methods.*

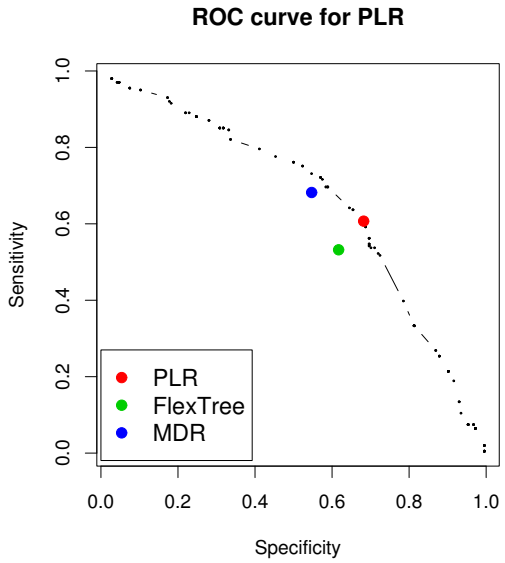

Figure 5: *Receiver operating characteristic (ROC) curve for penalized logistic regression. Penalized logistic regression achieved higher sensitivity and specificity than Flextree.*

As done in Section 5.1.1, we generated the receiver operating characteristic curve (Figure 5) for penalized logistic regression by varying the classification threshold between 0 and 1. Both sensitivity and specificity of Flextree are lower than those of penalized logistic regression; therefore, penalized logistic regression would achieve higher sensitivity (specificity)

than Flextree, if its specificity (sensitivity) is adjusted to be the same as Flextree. Although the sensitivity of MDR is higher than that of penalized logistic regression, the blue dot is still off the ROC curve, toward the lower left corner. In addition, sensitivity and specificity are more even for penalized logistic regression.

5.2.2 Bootstrap analysis of the feature selection

When we fit a penalized logistic regression model with forward stepwise selection using this bladder cancer dataset, the four terms in Table 10 were selected. To validate their significance, we performed a similar bootstrap analysis as in Section 5.1.2. The second column of Table 10 records the number of bootstrap runs (out of $B = 300$) in which the factors were chosen.

Factors selected from the whole data	Frequency
smoke status	296/300
MPO	187/300
GSTM1	133/300
GSTT1	128/300

Table 10: *Significant factors selected from the whole dataset (left column) and their frequencies in 300 bootstrap runs (right column).*

The factors/interactions of factors that were frequently selected through the bootstrap runs are listed in Table 11. The factors in Table 10 form the subset with the highest ranks among the ones listed in Table 11, providing evidence of reliability. The latter half of Table 11 shows that even the interaction terms with the largest counts were not as frequent as other common main effect terms. When we applied MDR, the second order interaction term *MPO* $\times$ *smoke status* was often selected; however, according to the bootstrap results, logistic regression method can explain their effect in a simpler, additive model. In addition, MDR was not able to identify other potentially important features.

Factor	Frequency	Interaction of factors	Frequency
smoke status	296/300	*smoke status* $\times$ *MPO*	38/300
MPO	187/300	*GSTT1* $\times$ *NAT2*	34/300
GSTM1	133/300	*GSTM1* $\times$ *MnSOD*	26/300
GSTT1	128/300	*GSTT1* $\times$ *XRCC1*	24/300
NAT2	88/300		
MnSOD	67/300		

Table 11: *Factors/interactions of factors that were selected in 300 bootstrap runs of forward stepwise procedure with relatively high frequencies.*

We also used the co-occurrence matrix of the factors in Table 11 as a dissimilarity measure and applied hierarchical clustering. One of the tightest clusters was the pair of *GSTM1* and

NAT2: they were in the model 133 and 88 times respectively, but coincided only 33 times, implying that *NAT2* was often used to replace *GSTM1*.

These results from the bootstrap analysis are consistent with the findings in Hung et al. (2004) in several ways. The factors with high frequencies (the first column of Table 11) are among the ones that were shown to be significantly increasing the risk of bladder cancer, through conventional analyses reported in Hung et al. (2004). Hung et al. also incorporated some known facts about the functional similarities of the genes and improved the estimates of the odds ratio. From this hierarchical modeling, *MPO*, *GSTM1*, and *MnSOD* achieved high odds ratios with improved accuracy. In addition, their analysis of gene-environment interaction showed that although smoking status itself was a significant factor, none of its interaction with other genes was strikingly strong. Similarly, as can be seen from Table 11, our bootstrap runs did not detect any critical interaction effect.

As seen in Section 5.1.2 with the hypertension dataset, we again confirm that the forward stagewise procedure yields a similar set of significant factors to the forward stepwise method; thus, we could use the two methods interchangeably. Table 12 lists the factors/interactions of factors that were chosen most frequently in 300 bootstrap runs of forward stagewise procedure. The single factors that were selected more than 100/300 times are identical to those from the forward stepwise procedure as in Table 11. The two-way interaction terms were rarely selected, but the two listed in Table 12 forms a subset of those in Table 11.

Factor	Frequency	Interaction of factors	Frequency
smoke status	295/300	*smoke status* × *MPO*	43/300
MPO	179/300	*GSTM1* × *MnSOD*	34/300
GSTM1	119/300		
GSTT1	104/300		
MnSOD	60/300		

Table 12: *Factors/interactions of factors that were selected in 300 bootstrap runs of forward stagewise procedure with relatively high frequencies.*

6 Discussion

We have proposed using logistic regression with a penalization on the size of the L_2 norm of the coefficients. The penalty was imposed not only for the usual smoothing effect but also for convenient and sometimes necessary features that the quadratic penalization accompanied. In regular logistic regression models, a small sample size prohibits high-order interaction terms, and variables with constant zero entries are often not allowed. Because these situations are common in modeling gene-gene interactions, logistic regression is limited in its applications. However, the quadratic penalization scheme yields a stable fit, even with a large number of parameters, and automatically assigns zero to the coefficients of zero columns.

We modified the hierarchy rule of the forward stepwise procedure to allow the interaction terms to enter the model more easily. One strategy was to accept an interaction term as a candidate if either component was already in the model. If a strong interaction effect with negligible main effects is suspected, more flexible rules, such as accepting an interaction term even with no main effect terms, should be applied. However, the forward stepwise procedure selects variables in a greedy manner. A less greedy selection through L_1 regularization will allow the terms to enter the model more smoothly.

Logistic regression yields a reasonable prediction accuracy and identification of significant factors along with their interaction structures. We have shown that adding a quadratic penalization is a simple but powerful remedy that makes it possible to use logistic regression in building gene-gene interaction models.

Appendix

A Dimensionality of MDR

Here we present the details of how the simulations were done to evaluate the effective degrees of freedom of the MDR fits, as briefly discussed in Section 2.1.2. First, we review a standard scenario for comparing nested logistic regression models. Suppose we have n measurements on two three-level factors F_1 and F_2, and a binary (case/control) response Y generated *completely at random* — i.e. as a coin flip, totally independent of F_1 or F_2. We then fit two models for the probabilities p_{ij} of a *case* in cell i of F_1 and cell j of F_2:

1. $\mathbf{p}^0$: a constant model $p_{ij} = p_0$, which says the probability of a case is fixed and independent of the factors (the correct model).

2. $\mathbf{p}^1$: a second-order interaction logistic regression model, which allows for a separate probability p_{ij} of a case in each cell of the 3×3 table formed by the factors.

If y_ℓ is the observed binary response for observation ℓ, and the model probability is $p_\ell = p_{i_\ell, j_\ell}$, then the *deviance* measures the discrepancy between the data and the model:

$$\mathrm{Dev}(\mathbf{y}, \mathbf{p}) = -2 \sum_{\ell=1}^{n} \left[y_\ell \log(p_\ell) + (1 - y_\ell) \log(1 - p_\ell) \right]. \tag{16}$$

We now fit the two models separately, by minimizing the deviance above for each, yielding fitted models $\hat{\mathbf{p}}^0$ and $\hat{\mathbf{p}}^1$. In this case the *change in deviance*

$$\mathrm{Dev}(\hat{\mathbf{p}}^1, \hat{\mathbf{p}}^0) = \mathrm{Dev}(\mathbf{y}, \hat{\mathbf{p}}^0) - \mathrm{Dev}(\mathbf{y}, \hat{\mathbf{p}}^1) \tag{17}$$

measures the improvement in fit from using the richer model over the constant model. Since the smaller model is correct in this case, the bigger model is "fitting the noise." Likelihood theory tells us that as the sample size n gets large, the change in deviance has a χ^2_8 distribution with degrees-of-freedom equal to $8 = 9 - 1$, the difference in the number of parameters

in the two models. If we fit an additive logistic regression model for $\mathbf{p}^1$ instead, the change in deviance would have an asymptotic χ^2_4 distribution. Two important facts emerge from this preamble:

- The more parameters we fit, the larger the change in deviance from a null model, and the more we overfit the data.
- The degrees-of-freedom measures the average amount of overfitting; indeed, the degrees of freedom d is the *mean* of the χ^2_d distribution.

This analysis works for models fit in linear subspaces of the parameter space. However, we can generalize it in a natural way to assess more complex models, such as MDR.

In the scenario above, MDR would examine each of the 9 cells in the two-way table, and based on the training data responses, create its "one-dimensional" binary factor F_M with levels *high-risk* and *low-risk*. With this factor in hand, we could go ahead and fit a two-parameter model with probabilities of a case p_H and p_L in each of these groups. We could then fit this model to the data, yielding a fitted probability vector $\hat{\mathbf{p}}^M$, and compute the change in deviance $\mathrm{Dev}(\hat{\mathbf{p}}^M, \hat{\mathbf{p}}^0)$. Ordinarily for a single two-level factor fit to a null model, we would expect a χ^2_1 distribution. However, the two-level factor was not predetermined, but *fit to the data*. Hence we expect change of deviances bigger than predicted by a χ^2_1. The idea of the simulation is to fit these models many many times to null data, and estimate the effective degrees of freedom as the average change in the deviance (Hastie & Tibshirani 1990).

We used this simulation model to examine two aspects of the effective dimension of MDR:

- For a fixed set of K factors, the effective degrees-of-freedom cost for creating the binary factor F_M.
- The additional cost for searching among all possible sets of size K from a pool of p available factors.

In our experiments we varied both K and p. The results are summarized in Section 2.1.2 along with Figure 1 and Table 1.

B Two-Locus Modeling

Many researchers have suggested methods to categorize two-locus models for genetic diseases and to mathematically formulate the corresponding probabilities of influencing the disease status. The two-locus models are often divided into two classes: *heterogeneity models* for which a certain genotype causes the disease independently of the genotype on the other locus, and *epistatis models* for which the two genotypes are dependent. To represent the two distinct classes, the concept of *penetrance* is often used. *Penetrance* is a genetic term meaning the proportion of individuals with a disease-causing gene that actually show the symptoms of the disease.

Let A and B denote potential disease-causing genotypes on two different loci, and use the following notation to formulate different genetic models:

$$\begin{aligned} f_A &= P(\text{Have disease}|A), \\ f_B &= P(\text{Have disease}|B), \text{ and} \\ f_{A,B} &= P(\text{Have disease}|A, B). \end{aligned}$$

That is, f_A, f_B, and $f_{A,B}$ denote the penetrance, or the conditional probability of resulting in the disease, for the individuals carrying the genotypes A, B, and both A and B, respectively.

Vieland & Huang (2003) defined *heterogeneity* between two loci to be the relationship with the following *fundamental heterogeneity equation*:

$$f_{A,B} = f_A + f_B - (f_A \times f_B), \tag{18}$$

which is directly derived from the following more obvious representation of the independent effect of a pair of genotypes:

$$1 - f_{A,B} = (1 - f_A) \times (1 - f_B).$$

They referred to any other two-locus relationship for which (18) does not hold as *epistatic*. Neuman & Rice (1992) distinguished the heterogeneity models likewise. Risch (1990) also characterized *multiplicative* and *additive* two-locus models; the disease penetrance for carriers of both genotypes A and B was multiplicative or additive in the penetrance scores for single genotypes A and B. Risch considered the additive model to be a reasonable approximation of a heterogeneity model.

As we demonstrated in Section 3.4 and Section 4, logistic regression identifies the relationship among the active genes as either additive or having an interaction; however, this distinction is not equivalent to that of heterogeneity and the epistatic relationship described above. For example, *epistatic model III* in Section 4 has a probability distribution such that the conditional probabilities of disease are additive in log-odds; we expect an additive model when applying logistic regression. Although in genetics, heterogeneity models are often characterized by no interaction among loci in affecting the disease, the factors are not necessarily conceived to be additive in logistic regression. However, in the example illustrated in Section 4, the interaction effect in the heterogeneity model was not as critical as in other epistatic models, and logistic regression found an additive model in more than 50% of the repeats.

Acknowledgments

The authors thank Richard Olshen for valuable comments and for a version of data from the SAPPHIRe project, John Witte for the bladder cancer dataset, and Amir Najmi and Balasubramanian Narasimhan for the codes for Flextree. The authors are also grateful to Robert Tibshirani and other Hastie-Tibshirani Lab members for helpful discussions. Trevor Hastie was partially supported by grant DMS-0505676 from the National Science Foundation, and grant 2R01 CA 72028-07 from the National Institutes of Health.

References

Coffey, C., Hebert, P., Ritchie, M., Krumholz, H., Gaziano, J., Ridker, P., Brown, N., Vaughan, D. & Moore, J. (2004), 'An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: The importance of model validation', *BMC Bioinformatics* **5**, 49.

Efron, B. & Tibshirani, R. (1993), *An Introduction to the Bootstrap*, CHAPMAN & HALL/CRC, Boca Raton.

Friedman, J. (1991), 'Multivariate adaptive regression splines', *The Annals of Statistics* **19**, 1–67.

Gray, R. (1992), 'Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis', *Journal of the American Statistical Association* **87**, 942–951.

Hahn, L., Ritchie, M. & Moore, J. (2003), 'Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interaction', *Bioinformatics* **19**, 376–382.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, CHAPMAN & HALL/CRC, London.

Hoerl, A. E. & Kennard, R. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**, 55–67.

Huang, J., Lin, A., Narasimhan, B., Quertermous, T., Hsiung, C., Ho, L., Grove, J., Oliver, M., Ranade, K., Risch, N. & Olshen, R. (2004), 'Tree-structured supervised learning and the genetics of hypertension', *Proceedings of the National Academy of Sciences* **101**, 10529–10534.

Hung, R., Brennan, P., Malaveille, C., Porru, S., Donato, F., Boffetta, P. & Witte, J. (2004), 'Using hierarchical modeling in genetic association studies with multiple markers: Application to a case-control study of bladder cancer', *Cancer Epidemiology, Biomarkers & Prevention* **13**, 1013–1021.

Le Cessie, S. & Van Houwelingen, J. (1992), 'Ridge estimators in logistic regression', *Applied Statistics* **41**, 191–201.

Lee, A. & Silvapulle, M. (1988), 'Ridge estimation in logistic regression', *Communications in Statistics, Simulation and Computation* **17**, 1231–1257.

McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models*, CHAPMAN & HALL/CRC, Boca Raton.

Neuman, R. & Rice, J. (1992), 'Two-locus models of disease', *Genetic Epidemiology* **9**, 347–365.

Risch, N. (1990), 'Linkage strategies for genetically complex traits. i. multilocus models', *American Journal of Human Genetics* **46**, 222–228.

Ritchie, M., Hahn, L. & Moore, J. (2003), 'Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity', *Genetic Epidemiology* **24**, 150–157.

Ritchie, M., Hahn, L., Roodi, N., Bailey, L., Dupont, W., Parl, F. & Moore, J. (2001), 'Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer', *American Journal of Human Genetics* **69**, 138–147.

Vieland, V. & Huang, J. (2003), 'Two-locus heterogeneity cannot be distinguished from two-locus epistasis on the basis of affected-sib-pair data', *American Journal of Human Genetics* **73**, 223–232.

Yuan, M. & Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society. Series B* **68**, 49–67.

Zhu, J. & Hastie, T. (2004), 'Classification of gene microarrays by penalized logistic regression', *Biostatistics* **46**, 505–510.