Ranking genetic factors related to age-related macular degeneration by variable selection confidence sets

Chao Zheng[1], Michael Zhang[2], Davide Ferrari *[1], and Paul Baird[2]

[1]School of Mathematics and Statistics, University of Melbourne
[2]Centre for Eye Research Australia (CERA), University of Melbourne

December 18, 2015

Abstract

The widespread use of generalized linear models in case-control genetic studies has helped identify many disease-associated risk factors typically defined as DNA variants, or single nucleotide polymorphisms (SNPs). Up to now, most literature has focused on selecting a unique best subset of SNPs based on some statistical perspective. In the presence of pronounced noise, however, multiple biological paths are often found to be equally supported by a given dataset when dealing with complex genetic diseases. We address the ambiguity related to SNP selection by constructing a list of models – called variable selection confidence set (VSCS) – which contains the collection of all well-supported SNP combinations at a user-specified confidence level. The VSCS extends the familiar notion of confidence intervals in the variable selection setting and provides the practitioner with new tools aiding the variable selection activity beyond trusting a single model. Based on the VSCS, we consider natural graphical and numerical statistics measuring the inclusion importance of a SNP based on its frequency in the most parsimonious VSCS models. This work is motivated by available case-control genetic data on age-related macular degeneration, a widespread complex disease and leading cause of vision loss.

Keywords: Variable selection confidence set, likelihood ratio test, predictor ranking, case-control genotype data, age-related macular degeneration

*Richard Berry Building, University of Melbourne, Parkville, 3010, VIC, Australia. E-mail: dferrari@unimelb.edu.au

1 Introduction

Age-related macular degeneration (AMD) is a widespread condition and a leading cause of blindness and central vision loss among adults over age 50; this disease represents a global burden with estimated worldwide prevalence of 8.7% (Jonas, 2014). AMD is often referred to as a complex disease, since it is caused by the interaction of a number of genetic, environmental and lifestyle factors, many of which have not yet been discovered. The motivating dataset in this paper consists of case-control categorical measurements at 20 DNA loci called single nucleotide polymporphisms (SNPs) published through our previous work as part of the AMD Gene Consortium (Fritsche et al., 2013), an international collaboration involving 18 research groups. SNPs are substitutions of a single nucleotide (A = Adenine, T= Thymine, C=Cytosine, G=Guanine) at a specific position on the genome. An understanding of their role in disease is sought in order to better diagnose, predict disease progression and personalize treatment regimens for patients. The main motivation of this paper is to select SNPs that have been identified as being associated with AMD and develop a reliable ranking mechanism to judge the importance of individual SNPs and their combinations.

A wealth of methods from the statistical literature of model selection may be used to select a single combination of SNPs. Examples in genetics include classic information-theoretical criteria, Bayesian and frequentist sparsity-inducing penalization approaches (e.g., see (Ayers and Cordell, 2010; Frommlet et al., 2012; Li et al., 2014; Park and Hastie, 2008; Wen, 2015; Wu et al., 2009)). Although the usual goal of model selection is to find a single optimal model from some perspective, in the presence of pronounced noise, multiple models may be equally supported by a given dataset. Dealing with this model ambiguity can be challenging in complex diseases such as AMD, where several genetic factors are likely to affect disease occurrence through a complex network of interactions.

In our view, focusing exclusively on a single selected model could imply loss of information from different perspectives. First, alternative explanations of disease etiology are tossed away – although both scientifically plausible and compatible with the data. Second, it is well known that the usual standard errors for a given selected model (e.g. regression model) fail to describe the variable selection uncertainty, so they cannot be directly used for establishing if a SNP subset is superior to other subsets. For similar reasons, ranking importance of SNPs based on the size of estimated coefficients

and their standard errors (e.g. z-scores) may be misleading. Although one might argue that ranking can be achieved by marginal measures of association (e.g. chi-square or exact Fisher tests), this would ultimately fail to capture the role of SNP combinations, which is a key goal when studying complex diseases.

This paper proposes to resolve the above issues related to model ambiguity by constructing a set of plausible SNP combinations, which we call a variable selection confidence set (VSCS). We begin by taking p SNPs where p is typically much smaller than the sample size, n. For instance, genome-wide scans and scientific considerations currently suggest that 20 genetic loci could play a role in AMD (Fritsche et al., 2013). Then, within a GLM framework, we compare all the models nested in the full model with p predictors using a likelihood ratio test (LRT) at a given significance level α. The final VSCS is then constructed by retaining the models not rejected by the LRT screening procedure.

An important consequence of the above LRT screening is that the resulting VSCS is guaranteed to contain the "true model" based on the available data with a probability of approximately $1 - \alpha$ (2.2). By analogy with the usual confidence intervals for parameter estimation, a VSCS should be regarded as a set of equally plausible models at the $100(1 - \alpha)\%$ given confidence level. Under standard conditions required to ensure the asymptotic distribution of the LRT screening for GLMs, the VSCS models tend to include the terms in the true model with large probability in large samples. Hence, the frequency of the selected SNPs and their combination in the VSCS models is expected to reflect their importance in relation to disease in a principled way that goes beyond just trusting a single model selected by some rule. In this paper, we use this property to obtain stable ranks for SNPs combinations in relation to AMD, and find stable "central" models by combining SNP predictors appearing with relatively high frequency in the most parsimonious VSCS models – the so-called lower boundary models (2.4).

The idea of frequentist confidence sets for variable selection has been recently explored by Ferrari and Yang (2015) in the context of linear regression models. Here, we extend that approach in the context of GLMs to analyze the AMD case-control genotype data. Previously, Hansen et al. (2011) proposed model confidence sets building upon step-down methods for multiple hypothesis testing (e.g., see Lehmann and Romano (2006)). Differently from their approach we focus specifically on LRT for GLMs in order to approximate coverage probability for the globally optimal model, which

is difficult to obtain when starting from an arbitrary list of models as in Hansen et al. (2011).

The rest of the paper is organized as follows. In Section 2, we present the main VSCS methodology for generalized linear models, discuss the notion of lower boundary models (LBMs) and introduce natural statistics based on the VSCS to rank predictors' inclusion importance. In the same section, we propose a model-combining strategy to obtain a single representative model based on the inclusion importance ranks of predictors. In Section 3, we study the finite sample properties of our methods using simulated genotype data. In Section 4, we apply the new methodology and study the AMD Gene Consortium case-control genotype data. In Section 5, we conclude and give final remarks.

2 Methods

2.1 Setup: Generalized linear models

Generalized linear models (GLMs) play an important role in many fields of empirical research. In genetics, due to their flexibility in modelling the relationship between predictors and a function of a response variable, GLMs have become popular tools to investigate the association between SNPs and phenotype (e.g. disease occurrence). Let Y_i, $i = 1, \ldots, n$, be independent phenotype measurements which are assumed to follow a distribution from an exponential family with mean $\mu_i = E(Y_i)$. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, $i = 1, \ldots, n$, be p-dimensional vectors of SNP covariates. Specifically, $x_{ij} = AA, Aa$ or aa, where the letters "A" and "a" represent one of the nucleotides in $\{A, T, C, G\}$. Other covariates representing demographic attributes of patients (e.g. gender, ethnicity, etc.) are collected in q-dimensional vectors, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iq})^T$, $i = 1, \ldots, n$.

Let $g(\cdot)$ be an invertible linearizing link function mapping the expectation of the response variable to the predictors as follows

$$g(\mu_i) = \eta_i = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i + \boldsymbol{\gamma}^T \mathbf{z}_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\beta_0 \in \mathbb{R}^1$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^T \in \mathbb{R}^q$ are model parameters. The intercept β_0 and the term $\boldsymbol{\gamma}^T \mathbf{z}_i$ are always included in the linear predictor (1), so the main focus is on selecting subsets of SNP predictors. Particularly, we assume that some of the coefficients in

$(\beta_1, \ldots, \beta_p)$ are zero and denote by $m^* \subseteq \{1, \ldots, p\}$ the set of indexes of all non-zero terms in the true model. The full model containing all p predictors is denoted by m_f and the set of all feasible models denoted by $\mathcal{M}$ is defined by taking all the nested models in m_f. We assume that the sample size n is sufficiently large so that estimates $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ can be obtained by standard likelihood methods. Alternatively, some conservative screening procedure may be used to reduce the number SNPs.

The methods discussed in the rest of this section are rather general and apply to any GLM as long as the conditions ensuring asymptotic normality of $(\hat{\alpha}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ are met. However, due the nature of the motivating AMD data application in this paper, both numerical experiments and real data analysis in Sections 3 and 4 will focus on the logistic regression model where Y_i is a binary response representing presence/absence of AMD and $E(Y_i) = P(Y_i = 1) = \mu_i$ represents the probability of disease for the ith subject. The relationship between response and covariates is modeled by the usual logit link function $\log(\mu_i) - \log(1 - \mu_i) = \eta_i$, where η_i is the linear predictor (1).

2.2 VSCS construction by likelihood ratio testing

Given observations $\{(Y_i, \mathbf{x}_i, \mathbf{z}_i), i = 1, \ldots, n\}$, our main interest is to construct a set of models, $\widehat{\mathcal{M}}_\alpha$, approximately satisfying $P(m^* \in \widehat{\mathcal{M}}_\alpha) \geq 1 - \alpha$, where $0 < \alpha < 1$ is a user-defined constant and m^* is the true model. The set $\widehat{\mathcal{M}}_\alpha$ is called a variable selection confidence set and all the models in $\widehat{\mathcal{M}}_\alpha$ may be regarded as equally plausible at the user-specified $100(1 - \alpha)\%$ confidence level (e.g., 95 or 99%) (Ferrari and Yang, 2015). To obtain $\widehat{\mathcal{M}}_\alpha$, we compare the full model m_f to a candidate model m nested in m_f by the likelihood ratio test (LRT) statistic

$$D_n(m) = 2\left\{\ell_n(m) - \ell_n(m_f)\right\}, \tag{2}$$

where $\ell_n(m)$ is the log-likelihood function evaluated at the maximum likelihood estimates for model m. Under the null hypothesis that the model m contains the true model, D_n follows a central chi-square distribution with $p - p_m$ degrees of freedom for large n.

The candidate model m survives the LRT evaluation if $D_n(m) < \chi^2(\alpha; p - p_m)$, where $\chi^2(\alpha; \nu)$ denotes the upper α-quantile for a chi-squared distribution with ν degrees of freedom, and p and p_m are the number of SNP predictors in m_f and m, respectively. Then we define the $(1 - \alpha)\%$-VSCS by

all the models surviving the LRT evaluation:

$$\widehat{\mathcal{M}}_\alpha = \left\{ m \in \mathcal{M} : D_n(m) < \chi^2(\alpha, p - p_m) \right\}. \tag{3}$$

The full model m_f is included in $\widehat{\mathcal{M}}$ by default. If a model does not survive the LRT evaluation, then such a model is considered overly risky, in the sense that it might miss relevant predictors at the confidence level $100(1-\alpha)\%$.

A direct consequence of this procedure is that the resulting VSCS $\widehat{\mathcal{M}}_\alpha$ includes the true model with large probability as $n \to \infty$. Specifically, if $m_f \neq m^*$ the VSCS has the approximate coverage probability

$$\lim_{n\to\infty} P\left(m^* \in \widehat{\mathcal{M}}_\alpha\right) \geq 1 - \alpha, \tag{4}$$

and in the special case where $m_f = m^*$, we have $\lim_{n\to\infty} P(m^* \in \widehat{\mathcal{M}}_\alpha) = 1$. This property follows directly from the well-known convergence in distribution of the likelihood ratio statistic D_n to the central chi-square distribution with $p - p_m$ degrees of freedom under the null hypothesis that the smaller candidate models are the true models.

The motivating data analysis problem in this paper allows us to compute the VSCS by an exhaustive search since the number of SNPs in the full model is relatively small ($p = 20$). In other applications, however, the VSCS might be quite large without further restrictions on the true regression terms, $\beta_1, \ldots, \beta_p$. In such situations, it is natural to resort to stochastic sampling techniques to draw models in $\widehat{\mathcal{M}}_\alpha$, or impose more stringent sparsity assumptions on the true model structure.

2.3 The lower boundary models

Without additional assumptions on the true model, the VSCS can be large because many models – roughly $2^{p-p_{m^*}}$ – containing the true model plus unimportant terms survive the LRT screening. We address the potential largeness of the VSCSs by focusing on a smaller but very informative subset of the VSCS – the set of lower boundary models (LMBs) – hereafter denoted by $\widehat{\mathcal{B}}_\alpha$. The lower boundary model set, $\widehat{\mathcal{B}}_\alpha$, is defined by all VSCS models that do not have nested sub-models in the VSCS. In lay terms, the LBMs can be regarded as a subset of maximally parsimonious models which are at the same time well-supported by the data.

Ferrari and Yang (2015)'s study of the LBMs in the context of linear models shows that both the cardinality of $\widehat{\mathcal{B}}_\alpha$ and the composition of its constituent models carry useful information about the overall variable selection uncertainty. It is easy to check that their results on the LBMs immediately hold for the GLM regression framework, when the number of predictors, p, is fixed and n is large. For the purposes of the current analysis, it is useful to summarize their results by distinguishing the following scenarios concerning the cardinality of $\widehat{\mathcal{B}}_\alpha$:

(i) *Zero variable selection uncertainty.* In the ideal case where we have overwhelming information in the data, $\widehat{\mathcal{B}}_\alpha = \{m^*\}$ with large probability as $n \to \infty$ (i.e. the LBM set contains only the true model). This represents the ideal situation where the sample provides us with enough information to detect all the predictors in m^* as they appear in the unique most parsimonious description of the data.

(ii) *Moderate variable selection uncertainty.* A more frequent situation in real applications occurs when $\widehat{\mathcal{B}}_\alpha$ contains more than 1 model, but its cardinality is much smaller than the entire VSCS, even in the presence of pronounced noise. If the models in $\widehat{\mathcal{B}}_\alpha$ differ by only a few predictors, the ones appearing with relatively large frequency should be regarded as important since they play a role in a number of parsimonious and well-supported explanations of the data.

(iii) *Strong variable selection uncertainty.* The most challenging case is when there is too much noise compared to signal in the sample. In this situation the size of $\widehat{\mathcal{B}}_\alpha$ can be very large and the individual constituents of $\widehat{\mathcal{B}}_\alpha$ may contain a small number of predictors. This means that predictor combinations are more or less picked at random by any model-selection method. In the least favorable case, the cardinality of $\widehat{\mathcal{B}}_\alpha$ can be as large as $\binom{p}{\lfloor p/2 \rfloor}$.

The above behavior of the LBMs is confirmed in the numerical experiments presented in Section 3, suggesting that the cardinality of $\widehat{\mathcal{B}}_\alpha$ and the composition of its constituent models are helpful indicators of the total variable selection uncertainty in a sample. As an illustration, Figure 1 (top panels) shows the number of models in $\widehat{\mathcal{M}}_\alpha$ (on the log-scale) and $\widehat{\mathcal{B}}_\alpha$ for different sample sizes. The graphs are based on genotype data simulated from a logistic regression model (Model 1 in Section 3 with $(p, \rho) = (10, 0)$ and m^* containing 5 predictors). Regardless of the confidence level, when n is relatively small there are multiple LBMs quite similar to the true model, each partly overlapping

with the true model. When n is large, $\widehat{\mathcal{B}}_\alpha$ tends to converge to a single model containing only the true predictors. Figure 1 (bottom panels) shows the average Hamming distance between the models in $\widehat{\mathcal{M}}_\alpha$ and $\widehat{\mathcal{B}}_\alpha$ and the true model m^*. The Hamming distance $d_H(m, m^*)$ between individual models in m and m^* is defined as the number of different terms in m compared to m^*. As n increases, the only remaining LBM is the true model.

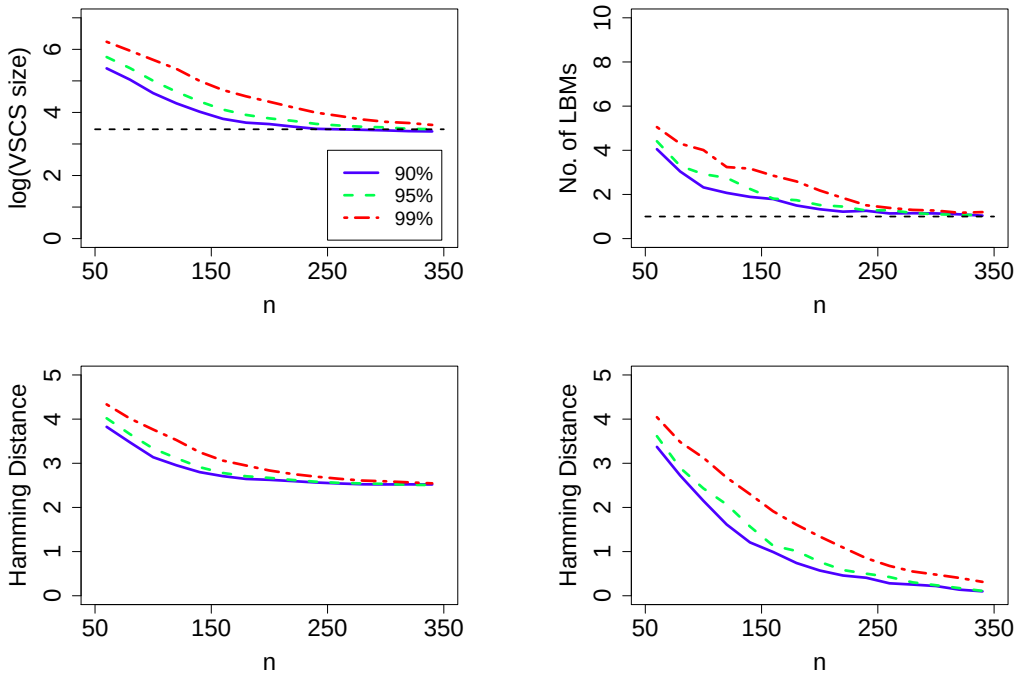

Figure 1: Top panels: Variable selection confidence set ($\widehat{\mathcal{M}}_\alpha$) and lower boundary model set ($\widehat{\mathcal{B}}_\alpha$) for increasing sample size (n) and different confidence levels (90, 95 and 99%) for logistic regression. Top panels: Cardinality of $\widehat{\mathcal{M}}_\alpha$ (on the log-scale) and $\widehat{\mathcal{B}}_\alpha$. Bottom panels: Average Hamming distance between all the models in $\widehat{\mathcal{M}}_\alpha$ and $\widehat{\mathcal{B}}_\alpha$ and the true model m^*. Each point in the curves is obtained by Monte Carlo averages based on 100 simulations from Model 1 described in Section 3 with $(p, \rho) = (10, 0)$.

2.4 Inclusion importance statistics

For linear models, Ferrari and Yang (2015) show that in large samples true regression terms tend to appear in the LBMs with probability near 1, while unimportant terms appear with small probability. Our Monte Carlo simulations in Section 3 confirm this understanding in the context of logistic regression and motivate the following statistics to rank the importance of predictors. The inclusion

importance (II) for the jth predictor is defined as

$$\widehat{II}_\alpha(j) = \frac{1}{|\widehat{\mathcal{B}}_\alpha|} \sum_{m \in \widehat{\mathcal{B}}_\alpha} I(x_j \in m), \tag{5}$$

where $\{x_j \in m\}$ denotes the event that the jth predictor is included in model m, $I(\cdot)$ is the indicator function and $|A|$ denotes the cardinality of the set A. If the jth predictor appears in all the boundary models then its inclusion importance is $\widehat{II}_\alpha(j) = 1$. If a predictor appears only in a few LBMs, its II value is near zero.

It is useful to generalize the idea of inclusion importance by looking at the joint and conditional frequencies for two or more predictors in $\widehat{\mathcal{B}}_\alpha$. The joint importance, or co-importance, of predictors j and k is defined by

$$\widehat{II}_\alpha(j,k) = \frac{1}{|\widehat{\mathcal{B}}_\alpha|} \sum_{m \in \widehat{\mathcal{B}}_\alpha} I(\{x_j \in m\} \cap \{x_k \in m\}). \tag{6}$$

We remark that this measure provides information on the joint utility of predictors in relation to the response in a way that goes beyond simply measuring association between predictors. For example, Figure 3 compares co-inclusion importance with pairwise sample mutual information (MI) for the SNPs in the AMD Genotype data detailed in Section 4. The sample MI measures the overlap between the distribution of two SNPs and is therefore a generic measure of association. Specifically, $MI = E_{\hat{p}_{jk}} \log\{\hat{p}_{jk}/(\hat{p}_j \hat{p}_k)\}$, where $\hat{p}_{jk}$ and $\hat{p}_j$, $\hat{p}_k$ denote estimated joint and marginal distributions for SNPs j and k. Note that while all but two SNP pairs are largely independent, several SNPs co-appear with others in the LBMs with respectable frequency. This means that such pairs contribute to a number of well-supported explanations of the response.

If $\widehat{II}_\alpha(k) > 0$, the conditional importance of predictors j given k is defined by

$$\widehat{II}_\alpha(j|k) = \frac{\widehat{II}_\alpha(j,k)}{\widehat{II}_\alpha(k)}. \tag{7}$$

If $\widehat{II}_\alpha(k) = 0$, predictor k does not appear in the LBMs and the conditional inclusion importance remains undefined. Given that predictor k is in the LBM, the conditional inclusion importance

statistic (7) can be used to summarize the conditional dependencies between k and predictors in the LBM in terms of their role in explaining the response variable.

The marginal and conditional importance statistics can be displayed in a graph as shown in Figure 4 to represent what we call an inclusion-importance network. The nodes in the graph represent SNPs with size proportional to the marginal inclusion statistic, $\widehat{II}_\alpha(j)$. Any two nodes, say j and k, are joined by two directed edges with thickness proportional to $\widehat{II}_\alpha(j|k)$ and $\widehat{II}_\alpha(k|j)$ with arrows pointing in the direction $k \to j$ and $j \to k$, respectively. Disconnected (or weakly connected) nodes represent conditionally independent (or weakly dependent) predictors, meaning that the inclusion of one predictor is not related to the importance of the others. Clearly, the overall degree of connectedness and centrality of a predictor in the graph summarize the importance of such a predictor in relation to others for explaining the response variable.

2.5 Combining lower boundary models

The set of lower boundary models is a summary statistic of the variable selection uncertainty in the sample, where each LBM offers a partial but very plausible view of the underlying data-generating process. It is now well understood that estimation or prediction can be improved by combining models (e.g. see Claeskens and Hjort (2008) for a book-length exposition). Thus, in the same spirit of model combining methods, we propose to aggregate the models in the set lower boundary models to achieve a better predictive performance than could be obtained from any individual constituent.

Suppose that $\widehat{\mathcal{B}}_\alpha$ contains $k \leq p$ distinct predictors and let $(j_1, \ldots, j_k)$ denote an arrangement of indexes in $\{1, \ldots, k\}$. The LBM-aggregated model, $\widehat{m}_{ag}$, is defined by the index set

$$\widehat{m}_{ag} = \left\{ (j_1, \ldots, j_{\widetilde{k}}) \in \widehat{\mathcal{M}}_\alpha : \widehat{II}_\alpha(j_1) \geq \widehat{II}_\alpha(j_2) \geq \cdots \geq \widehat{II}_\alpha(j_{\widetilde{k}}),\ \widetilde{k} \leq k \right\}, \tag{8}$$

where $\widehat{II}_\alpha(\cdot)$ is the statistic defined in (5). Then, the combined model $\widehat{m}_{ag}$ contains the most useful $\widetilde{k}$ predictors as measured by their marginal II values. Where appropriate, a value of $\widetilde{k}$ strictly smaller than k may be used to control the complexity of the final model and avoid predictors with nonzero but very low II value.

Although we do not offer a universally optimal rule for selecting $\widetilde{k}$, our numerical simulations sug-

gest that the following step-wise strategy works well in practice. Starting from $\widetilde{k} = 1$ we progressively add candidate terms based on their inclusion importance rank, $\widehat{II}_\alpha(j)$, and assess the resulting model using some model-selection criterion (e.g. AIC, BIC, etc). Then we stop if the given model-selection criterion cannot be improved and the resulting model is in $\widehat{\mathcal{M}}_\alpha$. In our Monte Carlo simulations and real data analysis (Sections 3 and 4), we consider AIC and BIC as model selection criteria, and refer to the resulting models as AIC-$\widehat{\mathcal{B}}_\alpha$ and BIC-$\widehat{\mathcal{B}}_\alpha$ models, respectively. Other selection criteria (e.g. cross-validation) may be used instead, but they are not explored in the current paper. Our numerical findings suggest that this strategy yields parsimonious yet very informative models typically outperforming popular model-selection methods that ignore model-selection uncertainty.

3 Monte Carlo simulations

The main aims of our simulation experiments are: 1) To investigate the finite sample coverage probability of the VSCS; 2) To study the cardinality of the VSCS and lower boundary model set in relation to model selection variability; and 3) To study the performance of the LBM-aggregation strategy described in Section 2.5.

Similarly to Pan (2009) and Han and Pan (2012), we simulate genotype data by a latent variable approach. We draw independent p-variate vectors, $\mathbf{u}_i = (u_{i1}, \ldots, u_{ip})^T$, $i = 1, \ldots, n$, from a multivariate normal distribution with zero mean vector, unit variance vector and covariance matrix denoted by $\boldsymbol{\Sigma}$. Then we create corresponding genotype vectors, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, with jth element defined as

$$x_{ij} = \begin{cases} AA, & u_{ij} < c_1, \\ aA, & c_1 \leq u_{ij} < c_2, \\ aa, & u_{ij} \geq ca_2, \end{cases}$$

for some constants c_1 and c_2, with AA, aA, aa denoting genotype labels. In the following simulations, we set $a_1 = \Phi^{-1}(1/3)$ and $\Phi^{-1}(2/3)$ where $\Phi(\cdot)$ is the standard normal cdf. For each genotype vector, $\mathbf{x}_i$, we then generate binary responses with probability $\pi_i = \exp(\eta_i)/(1 + \exp\{\eta_i\})$, where η_i is the linear predictor defined in (1) with $\boldsymbol{\gamma} = (0, \ldots, 0)^T$ and $\beta_0 = 0$. The following settings for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are considered:

Model 1: The first $k = p/2$ coefficients have the same size and the rest are equal to 0: $\beta_j = (-1)^j$, $j = 1, \ldots, k$, and $\beta_j = 0$, $j = k+1, \ldots, p$. The elements of $\mathbf{\Sigma}$ follow the Toeplitz structure $\mathbf{\Sigma}_{ij} = \rho^{|i-j|}$, $0 \leq \rho < 1$.

Model 2: Same as Model 1, but the correlation matrix $\mathbf{\Sigma}$ is a block diagonal matrix; each block has size $(p/4) \times (p/4)$. Elements within each block have correlations all equal to $0 \leq \rho < 1$, whilst between-block correlations equal to 0.

Model 3: The first $k = p/2$ coefficients have decreasing size and the rest are equal to 0: $\beta_j = (-1)^j/j, j = 1, \ldots, k$ and $\beta_j = 0, j = k+1, ..., p$. Toeplitz covariance structure as in Model 1.

Model 4: Decreasing coefficients as in Model 3 and block-covariance structure as in Model 2.

MC Experiment 1: Cardinality of $\widehat{\mathcal{M}}_\alpha$, $\widehat{\mathcal{B}}_\alpha$ and coverage. Tables 1 and 2 show Monte Carlo estimates for the coverage probability, $P(m^* \in \widehat{\mathcal{M}}_\alpha)$ (the probability that the VSCS includes true model), the cardinality of VSCS and LBM, $|\widehat{\mathcal{M}}_\alpha|$ and $|\widehat{\mathcal{B}}_\alpha|$, and the average number of predictors in the lower boundary models, $avg|\widehat{\mathcal{B}}_\alpha|$, under different values for the sample size (n), number of predictors (p), correlation between predictors (ρ), and significance level (α).

For all considered models, the estimated coverage probability is fairly close to the nominal significance level, α. As expected, as α decreases, the cardinality of both VSCS and LBMs increases. When the signal relative to noise is large – which occurs for example when the size of the coefficients, β_j, decreases, or the sample size, n, is small – we observe coverage probability slightly smaller than the nominal probability $1 - \alpha$. As the correlation between variates increases, the variable selection uncertainty increases, leading to larger cardinality for both VSCS and LBM. The coverage probability, however, is rather stable irrespective of the correlation structure used to generate the covariates.

It is important to note that, differently from the VSCS, the cardinality of the LBM is much smaller compared to the total number of feasible models, 2^p. At the same time, the LBMs contain similar information to the whole VSCS in terms of finding the nonzero regression terms (see discussion in Section 2.3 and MC Example 2 in this section). This confirms that the LBMs are informative summaries for measuring the variable selection uncertainty while their contained number allows in principle to develop efficient algorithms for their discovery in larger problems.

Finally, from Tables 1 and 2 one can see that the average size of the individual lower boundary

models is usually smaller than the size of the true model. Due to the effect of the variable selection uncertainty, each LMB contains a different subset of the true model m^*. Additional simulations show that is quite unlikely that a LBM contains unimportant variables. This confirms that the model combining strategy discussed in Section 2.5 represents a rather sound selection method in its own right.

MC Experiment 2: Model aggregation by importance ranking. In our second Monte Carlo experiment, we focus on the LBMs and study whether the occurrence of predictors in the LBM is informative on the true regression terms. To this end, we consider retaining the first $\widetilde{k}$ regression terms, ranked according to the inclusion importance statistic (5) introduced in Section 2.5. The number of terms appearing in the final selected model, $\widetilde{k}$, is chosen by step-wise selection using AIC and BIC scores. To assess the model-selection performance of our method, we compute Monte Carlo estimates of the Hamming distance $d_H(\hat{m}, m^*)$ defined by the number of different terms in a selected model $\hat{m}$ compared to the true model m^*.

Table 3 shows Monte Carlo estimates of the Hamming distance for our LBM-aggregation methods (AIC-$\widehat{\mathcal{B}}_\alpha$ and BIC-$\widehat{\mathcal{B}}_\alpha$) and classic forward-selection based on AIC and BIC scores (F-AIC and F-BIC). In addition, we show results for other common selection methods based on penalization approaches: least absolute selection and shrinkage operator (LASSO), smoothed clipped absolute deviation (SCAD), and mini-max convex penalization (MCP) approaches (e.g. see Fan and Li (2001); Zhang (2010)). To compute LASSO, SCAD and MCP we used the R package `ncvreg` with tuning parameters chosen by 5-fold cross validation. The proposed LBM-aggregation methods generally outperform all the other selection procedures relying on a single model in most cases, regardless of the model structure and sample size. As the sample size increases, the accuracy of the aggregation strategy based on LBMs is found to be significantly improved compared to single-model selection methods.

4 Analysis of the AMD Genotype data

The AMD Genotype data analyzed in this section consist of measurements on patients from either outpatient clinics at the Royal Victorian Eye and Ear Hospital or through private ophthalmology

		Model 1								Model 3							
	$n=$	100				200				100				200			
	$p=$	8		12		8		12		8		12		8		12	
	$\rho=$	0	3/4	0	3/4	0	3/4	0	3/4	0	3/4	0	3/4	0	3/4	0	3/4
	α																
$P(m^* \in \widehat{\mathcal{M}}_\alpha)$	0.10	86.0	87.4	81.6	86.8	91.0	88.2	85.4	87.4	86.6	83.4	85.2	85.8	85.0	87.6	87.4	89.8
	0.05	92.0	92.2	87.8	93.0	99.0	94.6	91.0	94.6	92.2	92.0	91.2	92.2	98.8	94.4	94.0	94.6
	0.01	98.4	98.6	96.0	98.6	99.0	99.6	97.6	98.0	98.0	97.4	98.0	97.8	98.8	99.4	97.4	98.8
$\lvert\widehat{\mathcal{M}}_\alpha\rvert$	0.10	22.5	52.5	159.7	532.5	14.6	20.5	61.4	139.3	88.1	124.2	1396.7	2039.8	57.6	88.6	933.9	1531.9
	0.05	28.9	73.0	233.5	790.5	15.9	26.9	73.7	204.8	106.0	150.8	1732.6	2441.2	71.2	105.2	1167.4	1806.3
	0.01	48.9	122.9	466.9	1474.6	18.9	47.1	110.6	413.7	139.3	193.8	2416.6	3153.9	96.8	136.2	1593.3	2326.5
$\lvert\widehat{\mathcal{B}}_\alpha\rvert$	0.10	1.4	2.3	2.9	4.4	1.1	1.3	1.3	2.3	1.9	3.4	3.6	5.4	1.5	1.8	2.7	2.9
	0.05	1.6	2.8	3.6	4.9	1.1	1.6	1.5	2.8	2.1	4.1	4.1	6.6	1.6	2.0	2.6	3.4
	0.01	2.2	3.9	5.1	6.4	1.1	2.0	1.9	4.0	3.0	5.6	5.4	8.7	1.5	2.7	2.5	4.8
$avg(\widehat{\mathcal{B}}_\alpha)$	0.10	3.6	2.8	5.4	4.1	4.0	3.7	6.0	5.4	1.8	1.5	2.1	1.8	2.3	1.7	2.5	1.9
	0.05	3.3	2.4	5.0	3.6	3.9	3.5	5.8	5.0	1.5	1.3	1.8	1.5	2.0	1.5	2.2	1.7
	0.01	2.8	1.8	4.4	2.8	3.7	2.9	5.4	4.4	1.3	1.1	1.5	1.2	1.6	1.2	1.7	1.4

Table 1: Monte Carlo estimates of the coverage probability, $P(m^* \in \widehat{\mathcal{M}}_\alpha)$, cardinality of VSCS and LBM set, $|\widehat{\mathcal{M}}_\alpha|$ and $|\widehat{\mathcal{B}}_\alpha|$, and average number of predictors in the lower boundary models, $avg|\widehat{\mathcal{B}}_\alpha|$, under different values for the sample size (n), number of predictors (p), correlation between predictors (ρ), and significance level (α). Results are obtained from 500 Monte Carlo runs bases on Models 1 and 3 (constant and decreasing size of regression coefficients, respectively) with Toepliz covariance structure for the covariates. Monte Carlo standard errors are smaller than 0.1.

		Model 2				Model 4			
	$n =$	100		200		100		200	
	$p =$	8	12	8	12	8	12	8	12
	α								
$P(m^* \in \widehat{\mathcal{M}}_\alpha)$	0.10	85.0	83.2	88.2	86.8	89.6	84.8	89.4	89.6
	0.05	93.0	90.2	96.0	92.2	93.8	92.6	94.8	94.6
	0.01	98.8	98.0	99.4	98.4	98.2	97.6	99.0	99.2
$\lvert\widehat{\mathcal{M}}_\alpha\rvert$	0.10	38.7	430.1	17.7	110.6	128.3	2031.6	81.9	1454.1
	0.05	55.0	639.0	21.8	159.7	153.6	2428.9	98.3	1744.9
	0.01	99.3	1232.9	33.3	323.6	195.3	3108.9	132.6	2277.4
$\lvert\widehat{\mathcal{B}}_\alpha\rvert$	0.10	1.7	3.7	1.2	2.1	3.5	5.2	1.8	2.7
	0.05	2.1	4.4	1.2	2.4	4.3	6.1	2.1	3.0
	0.01	3.2	5.4	1.4	3.4	5.8	8.3	2.8	3.8
$avg(\widehat{\mathcal{B}}_\alpha)$	0.10	3.0	4.3	3.8	5.5	1.5	1.8	1.9	1.9
	0.05	2.6	3.9	3.6	5.1	1.3	1.5	1.6	1.7
	0.01	2.0	3.0	3.1	4.5	1.1	1.3	1.3	1.4

Table 2: Monte Carlo estimates of the coverage probability, $P(m^* \in \widehat{\mathcal{M}}_\alpha)$, cardinality of VSCS and LBM, $|\widehat{\mathcal{M}}_\alpha|$ and $|\widehat{\mathcal{B}}_\alpha|$ set, and average number of predictors in the lower boundary models, $avg|\widehat{\mathcal{B}}_\alpha|$, under different values for the sample size (n), number of predictors (p), correlation between predictors (ρ), and significance level (α). Results are obtained from 500 Monte Carlo runs based on Models 2 and 4 (constant and decreasing size of regression coefficients, respectively) with block-diagonal covariance structure ($\rho = 3/4$). The case of uncorrelated covariates for Models 2 and 4 is the same columns in Table 1 with $\rho = 0$ for Models 1 and 3. Monte Carlo standard errors are smaller than 0.1.

practices in Melbourne, Australia. Control subjects were selected from the same community. The subjects were Caucasian of Anglo-Celtic ethnic background. Patient collection and clinical examination were undertaken as described previously in Baird et al. (2004) . The cohort consisted of 418 subjects with advanced AMD and 266 healthy control subjects with no AMD. For each patient, demographic information was collected along with age and gender and included in our models to control for potential confounding. The 20 SNPs considered here include 14 well-replicated variants in the AMD risk assessment literature as well as 6 newly discovered variants (Fritsche et al., 2013). Genomic DNA was prepared from peripheral venous blood with genotyping of the 20 SNPs performed on the Mass Array platform (SEQUENOM, San Diego, CA) at the Murdoch Children's Research Institute, Melbourne as previously described in Gu et al. (2013).

n	p	ρ	$\alpha =$	AIC-$\widehat{\mathcal{B}}_\alpha$ 0.10	AIC-$\widehat{\mathcal{B}}_\alpha$ 0.05	AIC-$\widehat{\mathcal{B}}_\alpha$ 0.01	BIC-$\widehat{\mathcal{B}}_\alpha$ 0.10	BIC-$\widehat{\mathcal{B}}_\alpha$ 0.05	BIC-$\widehat{\mathcal{B}}_\alpha$ 0.01	F-AIC	F-BIC	LASSO	SCAD	MCP
100	8	0		0.78	0.90	1.17	0.79	0.95	1.20	3.06	3.01	1.51	1.31	2.16
		3/4		2.00	2.25	2.81	2.23	2.47	2.95	3.09	3.03	2.46	2.31	2.75
	12	0		1.56	1.59	1.64	1.62	1.61	1.63	3.20	3.55	1.99	1.83	3.17
		3/4		3.17	3.32	4.14	3.80	3.90	4.45	4.95	5.00	4.11	4.15	4.36
200	8	0		0.21	0.15	0.27	0.16	0.12	0.26	1.69	2.06	1.52	1.06	2.36
		3/4		0.83	1.06	1.49	0.91	1.05	1.50	1.53	1.85	1.52	1.32	2.65
	12	0		0.42	0.52	0.77	0.30	0.40	0.66	2.38	2.94	2.12	1.48	3.58
		3/4		1.49	1.68	1.82	1.59	1.89	1.98	3.44	3.79	2.38	2.00	3.91

Table 3: Monte Carlo estimates of the Hamming distance between the true model m^* and models selected by LBM-aggregation based on AIC and BIC (AIC-$\widehat{\mathcal{B}}_\alpha$ and BIC-$\widehat{\mathcal{B}}_\alpha$), forward-selection based on AIC and BIC (F-AIC and F-BIC), and LASSO, SCAD, MCP sparsity-inducing penalization approaches. For the LBM-aggregation method we consider 99.9, 99, 95 and 90% confidence levels. The tuning parameters for LASSO, SCAD and MCP methods are chosen by 5-fold cross validation. Results are based on 500 simulations from Model 1 under different values for the sample size (n), number of predictors (p) and correlation between predictors (ρ). Monte Carlo standard errors are smaller than 0.1.

VSCS, LBMs and ranks of SNPs. Table 4 shows summary statistics for the VSCS and LBM set at the 90, 95 and 99% confidence levels. As expected, the cardinality of both $\widehat{\mathcal{M}}_\alpha$ and $\widehat{\mathcal{B}}_\alpha$ decrease as α increases and the cardinality for the variable selection confidence set, $|\widehat{\mathcal{M}}_\alpha|$, is much larger compared to that of the LBM set, $|\widehat{\mathcal{B}}_\alpha|$. The number of lower boundary models is relatively large (43, 79 and 103) with a median number of SNPs for individual LBMs ranging from 6 to 8, thus evidencing considerable model uncertainty. We also report the average hamming distance (AHD) for all model pairs in the LBM computed as

$$AHD = \binom{|\widehat{\mathcal{B}}_\alpha|}{2}^{-1} \sum_{m,m' \in \widehat{\mathcal{B}}_\alpha : m \neq m'} d_H(m, m'),$$

where $d_H(m, m')$ denotes the number of SNPs at which m and m' are different. This quantity remains stable with α and is not negligible compared to the total number of SNP (20 SNPs), which indicates heterogeneous SNP combinations in the LBMs. Overall, the findings in Table 4 show that

the variable-selection uncertainty is quite strong, despite the relatively large sample size ($n = 684$) and suggest that the data may not be be well represented unequivocally by a single logistic regression model selected by a method that ignores the selection-uncertainty.

α	$\lvert\widehat{\mathcal{M}}_\alpha\rvert$	$\lvert\widehat{\mathcal{B}}_\alpha\rvert$	No. predictors in LBM					AHD
			min	Q1	median	Q3	max	
0.10	30827	43	6.0	7.0	7.0	8.5	10.0	5.2
0.05	57273	79	5.0	7.0	8.0	8.0	10.0	5.9
0.01	156459	103	4.0	5.0	6.0	7.0	10.0	6.1

Table 4: Variable selection summary statistics for the AMD Gene Consortium data: Cardinality of VSCS and LBM sets ($\lvert\widehat{\mathcal{M}}_\alpha\rvert$ and $\lvert\widehat{\mathcal{B}}_\alpha\rvert$), 5-number summary statistics for the number of predictors in the LMBs, and average Hamming distance (AHD) between pairs of LMB models, for different confidence levels ($\alpha = 0.1, 0.05, 0.01$).

Figure 2 summarizes SNP ranking according to the marginal inclusion importance statistic defined in (5). Note that 15 SNPs exhibit non-zero importance, about half of which receive II values greater than 1/2. Figure 4 illustrates the conditional inclusion importance graph for SNPs in the AMD data. A few SNPs with large marginal inclusion importance are central in the graph (e.g. see `rs4420638`, `rs3130783`). This suggests that such SNPs may play an important role in relation to AMD etiology in the presence of a number of other SNPs connected in the graph. On the other hand, certain SNPs, such as `rs10490924`, seem to be important in their own right regardless of the presence of other SNPs. Such SNPs lay on the periphery of the inclusion importance graph, meaning that they do not explain much of the disease occurrence conditionally to the presence of other SNPs.

The contention that removal of predictors that are essentially predictive of each other is common in practice under the assumption of linkage disequilbrium (LD) (i.e. nonzero association between SNPs). For example, SNPs `rs10490924` and `rs11200638` in the *ARMS2/HTRA1* region are highly correlated (Figure 3, right). However, our methodology suggests that the inclusion of either SNP does not automatically lead to the exclusion of the other. Actually, although `rs11200638` has large marginal II value compared to that of `rs10490924`, Figure 3 (left) shows that the two SNPs are important in explaining AMD occurrence when they are observed simultaneously, suggesting that there may be more unexplained factors at play.

In the conditional inclusion importance graph, of note is the centrality of the ApoE gene. Changes in this gene are known to be related to AMD. It is presently unknown what the role of ApoE is

in relation to the functional mechanism of effect with respect to allelic variation, although it is widely known that allelic variation can lead to significantly increased risk of risk development of the aforementioned diseases. Figure 3, suggest that multiple genetic variants together with ApoE are intertwined with disease etiology.

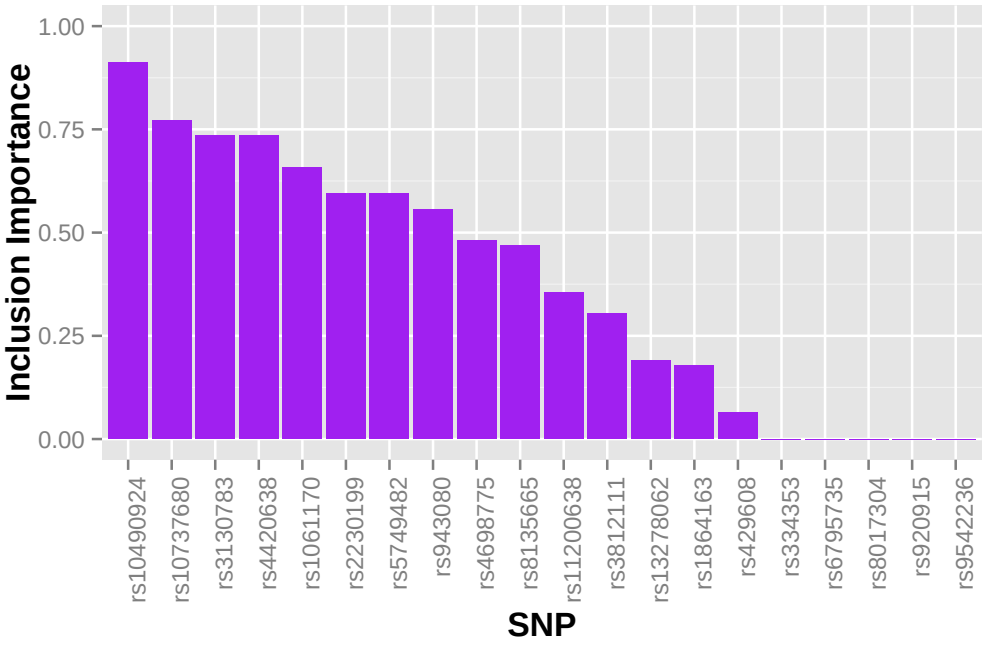

Figure 2: SNPs ranking for the AMD gene consortium data based on the marginal inclusion importance statistic, $\widehat{II}$, defined in (5). The inclusion importance statistic is computed by the frequency of the SNPs in the lower boundary model set $\widehat{\mathcal{B}}_\alpha$ as described in Section 2.4.

A comparison with previously reported and newly discovered loci. Columns 1–3 in Table 5 list newly discovered and established SNPs. Columns 4–5 compare the SNPs newly discovered by the AMD Gene Consortium with the results of our ranking method using the 95% confidence level. It is reassuring that our approach identifies most of the well-replicated SNP predictors, confirming the validity of our approach. Specifically, for the 12 well-replicated SNP variants, all except SNP `rs920915` show non-zero inclusion importance, largely confirming the current consortium list. In addition, our method finds two SNPs not currently part of the current consortium list (`rs1061170` and `rs11200638` with respectable II values ranking 5th and 11th). These are well known SNPs reputed important by other studies corresponding to the CFH and HTRA1 genes. Of the 6 newly found SNPs, only two receive non-zero inclusion importance, meaning that there is little evidence

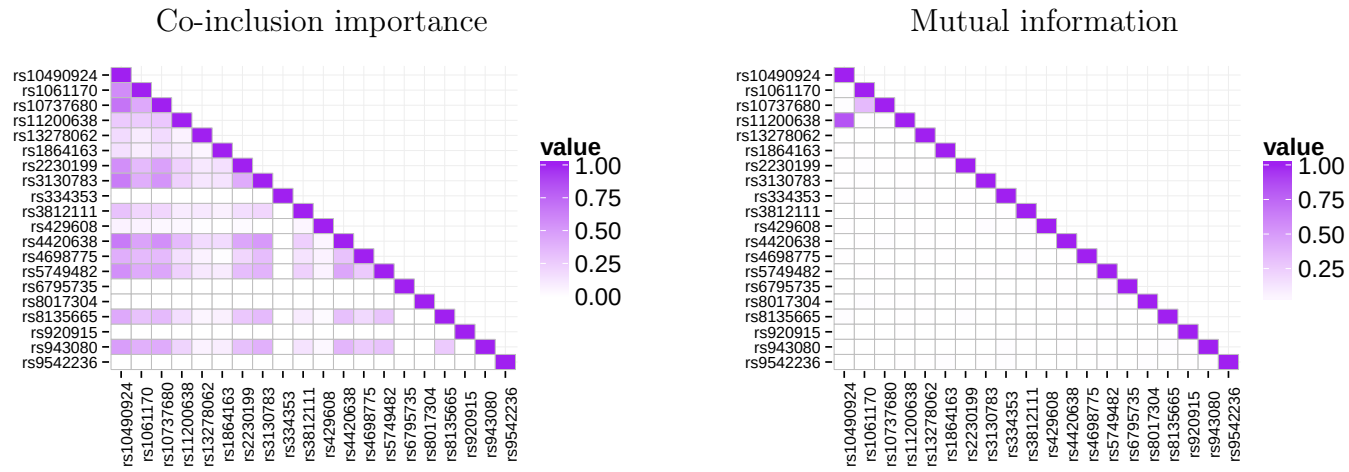

Figure 3: Co-inclusion importance at the 95% confidence level and sample mutual information for pairs of SNPs in the AMD genotype data. Left: Standardized co-inclusion importance $\widehat{II}_\alpha(j,k)/\widehat{II}_\alpha(j \text{ or } k)$, where $\widehat{II}_\alpha(j \text{ or } k) = \widehat{II}_\alpha(j) + \widehat{II}_\alpha(k) - \widehat{II}_\alpha(j,k)$ is the relative frequency of either SNP j or k in the LBMs. Right: Pair-wise sample mutual information for SNPs x_j and x_k.

in these data to suggest that the remaining 4 SNPs contribute much information to explain AMD occurrence beyond that already provided by the other SNPs with positive inclusion importance values.

Model combining and variable selection. It may be reasoned that the SNP predictors with highest II values are also the most pertinent to disease etiology. However, there is no total presence of any single SNP in every LBM model. The implication of a lack of unanimity points toward multiple possible disease pathways and highlights the variable selection uncertainty faced in generating AMD risk models. Nonetheless, a collection of stable SNPs can be obtained using the aggregation strategy proposed in Section 2.5.

Table 5 (Columns 6–9) shows the result of our LBM-aggregation approach via AIC and BIC dimension reduction (AIC-$\widehat{\mathcal{B}}_\alpha$ and BIC-$\widehat{\mathcal{B}}_\alpha$). For illustration purposes, we also report the selection obtained by other popular model-selection methods. First, note that none of the SNPs with zero II values are selected by the considered variable selection methods, confirming the stability of our ranking scheme. Second, we illustrate that some of the models selected by standard methods may miss some of the important SNPs at the 95% confidence level. Specifically, the LRT p-value for the BIC model selected by forward search (F-BIC) is smaller than $\alpha = 0.05$. This means that the F-BIC model is outside the VSCS and should be regarded as overly parsimonious since it is likely to miss one or more important SNPs related to AMD.

SNP id	Chromosome	Gene	Consortium	*II*	**AIC-$\widehat{\mathcal{B}}_\alpha$**	**BIC-$\widehat{\mathcal{B}}_\alpha$**	F-AIC	F-BIC	LASSO	SCAD	MCP
Loci previously reported loci with marginal p-values $< 5 \times 10^{-8}$:											
rs10490924	10	*ARMS2/HTRA1*	✓	0.91	✓	✓	✓	✓	✓	✓	✓
rs1061170	1	*CFH*		0.66	✓	✓	✓		✓		
rs10737680	1	*CFH*	✓	0.77	✓	✓	✓	✓	✓	✓	✓
rs11200638	10	*ARMS2/HTRA1*		0.35			✓				
rs13278062	8	*TNFRSF10A*	✓	0.19					✓	✓	✓
rs1864163	16	*CETP*	✓	0.18							
rs2230199	19	*C3*	✓	0.59	✓		✓		✓	✓	✓
rs3812111	6	*COL10A1*	✓	0.30					✓	✓	
rs429608	6	*C2/CFB*	✓	0.06					✓		
rs4420638	19	*APOE*	✓	0.73	✓	✓			✓	✓	✓
rs4698775	4	*CFI*	✓	0.48	✓		✓		✓	✓	✓
rs5749482	22	*TIMP3*	✓	0.59	✓				✓	✓	✓
rs920915	15	*LIPC*	✓	0.00					✓		
rs943080	6	*VEGFA*	✓	0.56	✓		✓		✓	✓	✓
Newly discovered loci with marginal p-values $< 5 \times 10^{-8}$:											
rs3130783	6	*IER3/DDR1*	✓	0.73	✓	✓	✓		✓	✓	✓
rs8135665	22	*SLC16A8*	✓	0.47					✓	✓	✓
rs334353	9	*TGFBR1*	✓	0.00							
rs8017304	14	*RAD51B*	✓	0.00					✓		
rs6795735	3	*ADAMTS9/MIR548A2*	✓	0.00					✓		
rs9542236	13	*B3GALTL*	✓	0.00					✓		
LRT p-value =					0.6913	0.0625	0.5692	0.00037	0.7950	0.2832	0.2583

Table 5: SNPs selection and ranking for the AMD gene consortium data grouped in previously reported (Rows 1–14) and newly discovered (rows 15–20) SNPs in the AMD literature. Columns 1–3: SNP, chromosome and gene identifiers. Column 4: SNPs previously identified as important by the AMD Gene Consortium. Column 5: Marginal inclusion importance statistic defined in (5). Columns 6–7: Models selected by the LBM-aggregation strategy described in Section 2.5 with number of variables, $\widetilde{k}$, selected via AIC and BIC (AIC-$\widehat{\mathcal{B}}_\alpha$ and BIC-$\widehat{\mathcal{B}}_\alpha$). Columns: 8–9: Models selected by AIC and BIC step-wise forward search. The last row denotes the p-value from the LRT comparing the full model with 20 predictors with the selected models.

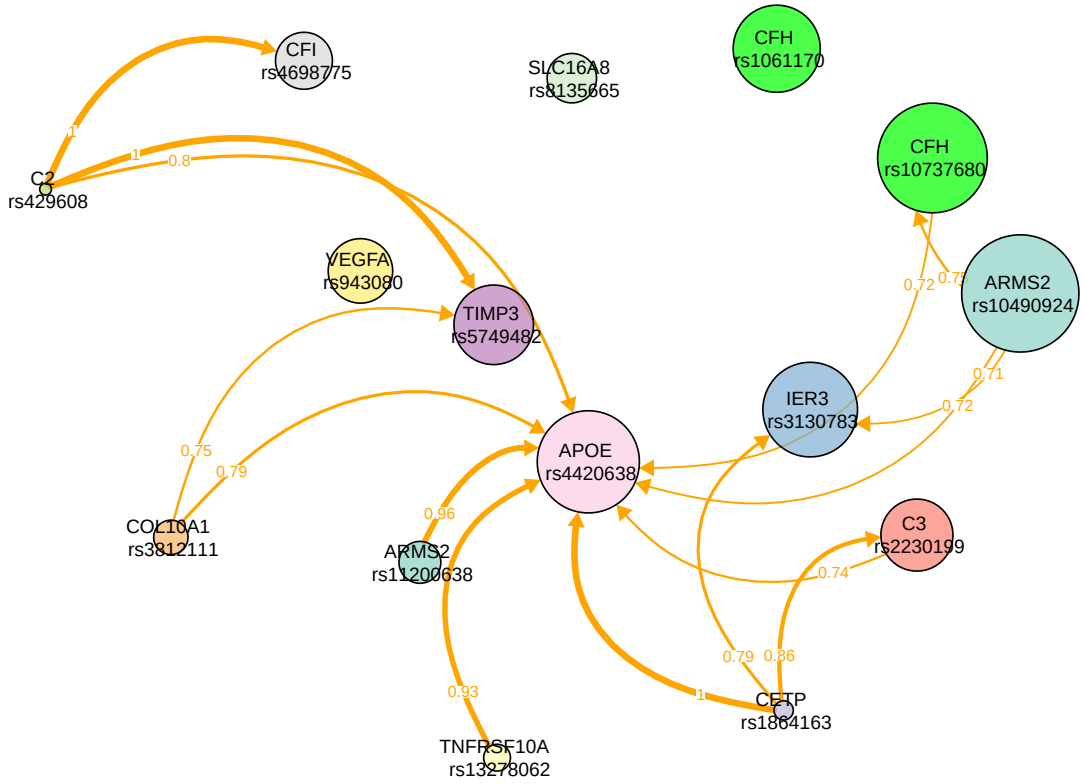

Figure 4: Conditional inclusion importance graph at the 95% confidence level for SNPs in the AMD genotype data. The size of the nodes and thickness of edges is proportional to marginal and conditional inclusion importance defined in (5) and (7), respectively. Edges with conditional inclusion importance less than 0.7 are omitted for clarity.

Finally, we use operating characteristic (ROC) curves to illustrate the performance of the selected logistic regression models. Figure 5 shows ROC curves representing the true positive rate against the false positive rate at various threshold settings for the classifier implied by logistic regression model; in the same figure, we report the area under the curve (AUC) for the considered methods. The ROC curves are obtained by 5-fold cross-validation with fixed models as listed in Table 5. Note that the performance of the AIC- and BIC-$\widehat{\mathcal{B}}_\alpha$ models is very close to that of the full model. The LBM-$\widehat{\mathcal{B}}_\alpha$ appears to improve the performance of the full model for false positives ranging from about 0.1 to 0.7, while the LBM-BIC does better than the full model for false positives $<$0.5 and then is close to the true model after that. This analysis confirms that the additional SNPs in the full model do not contain information helping reduce the mis-classification error.

Finally, note that the BIC model selected by forward search (F-BIC) is inferior to both LBM-

$\widehat{\mathcal{B}}_\alpha$ and LBM-$\widehat{\mathcal{B}}_\alpha$ models, confirming our previous understanding that this model may be overly parsimonious and misses some potentially useful SNPs. In additional analyses non presented here, we considered including two-way as well as higher order interactions in addition to the main effects selected by our method. However, assessment by AIC, BIC and cross-validation suggested that these more complex models are not superior to the AIC-$\widehat{\mathcal{B}}_\alpha$ and BIC-$\widehat{\mathcal{B}}_\alpha$ models containing only the main effects.

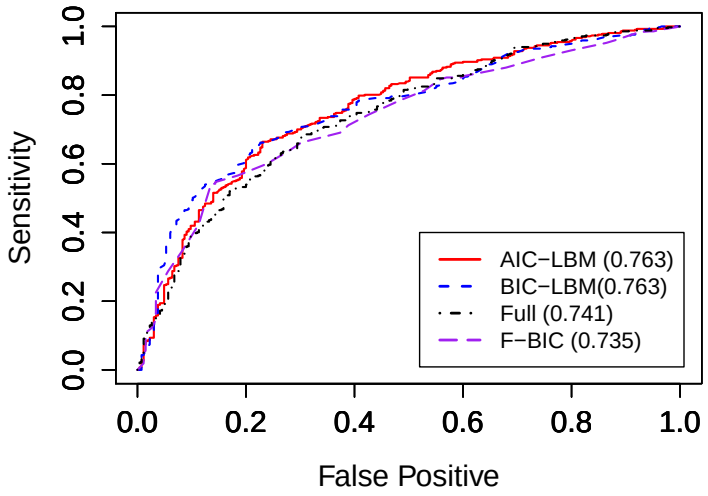

Figure 5: Out-of-sample receiver operating characteristic (ROC) curves for the AMD Gene Consortium data. The curves correspond to AIC- and BIC-LBM aggregation strategies, and the full model with 20 SNPs. Curves obtained using 5-fold cross-validation. For each model, we report in parenthesis the corresponding area under the curve (AUC).

5 Conclusion and final remarks

Variable selection uncertainty is a very common but often overlooked issue in applications of generalized linear models. In case-control genetic studies, the common practice of explaining a phenotype in terms of a single SNP combination may be inadequate when pronounced noise prevents one to learn whether the selected combination of predictors is the true explanation behind disease. As a consequence, often multiple SNP combinations are found to be compatible with a sample depending on the level of noise, statistical model adopted, and variable-selection method. To address this model-selection ambiguity, we extended previous work by Ferrari and Yang (2015) and introduced the notion of variable selection confidence set (VSCS) based on LRT screening for GLMs. We applied

the new approach to rank genetic factors underpinning AMD – an important eye disease with high prevalence in the global population – obtaining a stable selection of SNPs based on available AMD case-control data.

The VSCS contains models statistically equivalent to the true model at a pre-specified confidence level (e.g. 99% or 95%), thus representing a natural extension of the familiar confidence intervals for parameter estimation in the variable selection setting. We evaluated the validity of VSCSs constructed by LRT screening using simulated SNP data where the signal is not so clear, which we believe is very common in complex diseases. With adequate sample size, our results indicate that the coverage probability of the true model is near the nominal confidence level. Our numerical findings confirm earlier results by Ferrari and Yang (2015) in the context of linear regression and suggest that the cardinality and composition of the lower boundary models can be used to assess both variable selection uncertainty and learn the true model structure.

The VSCS methodology provides the practitioner with new tools in support of the variable selection activity in a way that goes beyond trusting a single selected model. Here we advocate the importance of a special subset of the VSCS, the lower boundary model set, $\widehat{\mathcal{B}}_\alpha$ (i.e. the set of maximally parsimonious models surviving the LRT screening). Our numerical examples confirm that, collectively, the models in $\widehat{\mathcal{B}}_\alpha$ contain a wealth of information on the true underlying model structure which can be used for variable ranking. Specifically, the frequency of predictors appearing in the LBMs (i.e. the II statistic defined in (5)) is a natural measure of their relevance in explaining the response. For example, a predictor with II value near 1 plays a role in most of the parsimonious and plausible explanations of the data. Ranking predictors based on the II statistics is found to be a reliable indicator of whether a given regression term is contained in the true model in presence of pronounced noise. Similarly, the joint (conditional) frequency of two or more predictors in the LMBs is informative on their joint (conditional) importance for explaining the response. In Sections 2.4 and 4 we explained how this information can be effectively conveyed visually, for example using common directed graphs.

Another contribution of this paper is a new approach for selecting a single central model based on the II ranks of predictors (Section 2.5). This strategy shares the strength of stability selection methods introduced first by Meinshausen and Bühlmann (2010), whereby a final stable model is

selected by taking predictors with sufficiently large frequency in a set of plausible models. While in stability selection the set of plausible models is obtained by re-sampling, here we use the LBMs obtained by LRT screening. Our Monte Carlo simulations suggest that model aggregation by II-ranking is a rather effective model-selection strategy in its own right typically outperforming other established model-selection approaches that do not benefit from selection-stability control. Due to these promising results, we believe that developing a theoretical understanding of the properties of this new aggregation approach in the future, including optimal selection of the reduced model size $(\widetilde{k})$, would be very valuable.

In the AMD Genotype data application, various models selected by certain classic variable selection methods are suboptimal in finding SNPs that are known to explain AMD etiology. Particularly, aggressive selection methods promoting very sparse models such as BIC are found to fall outside the VSCS. The usefulness of the proposed statistical framework for novel variant discovery is shown in the identification of SNPs in a set of well-replicated variants. The lower boundary set of models was found to identify the vast majority of well-replicated variants presented by the AMD Gene Consortium. In novel variant discovery, the VSCS aids by screening out newly found SNPs by genome-wide scans that do not contribute much information when considered together with well-established SNPs.

In the AMD Genotype data motivating this paper, the total number of SNPs in the full model is relatively small ($p = 20$). Thus, our methodology is adequately handled by generation of the entire confidence set and exhaustive LRT screening. Clearly, the exponential increase in models per additional SNP forces alternate methods more efficient at deriving a representative confidence and lower boundary sets. To address this issue, in our future work, we plan to extend the current procedure and develop a screening approach based on test statistic designed to handle a large number of predictors. These may be complemented by computationally efficient methods that can handle computation on larger model spaces. For example, we believe that certain Markov Chain Monte Carlo methods such as the reworked Gibbs sampler described in Qian and Field (2002) may be used to overcome these computational issues.

References

K. L. Ayers and H. J. Cordell. Snp selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology*, 34(8):879–891, 2010.

P. N. Baird, E. Guida, D. T. Chu, H. T. Vu, and R. H. Guymer. The $\varepsilon 2$ and $\varepsilon 4$ alleles of the apolipoprotein gene are associated with age-related macular degeneration. *Investigative Ophthalmology & Visual Science*, 45(5):1311–1315, 2004.

G. Claeskens and N. L. Hjort. *Model selection and model averaging*. Cambridge University Press, 2008.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

D. Ferrari and Y. Yang. Confidence sets for variable selection by f-testing. *Statistica sinica*, 2015.

L. G. Fritsche, W. Chen, M. Schu, B. L. Yaspan, Y. Yu, G. Thorleifsson, D. J. Zack, S. Arakawa, V. Cipriani, S. Ripke, et al. Seven new loci associated with age-related macular degeneration. *Nature Genetics*, 45(4):433–439, 2013.

F. Frommlet, F. Ruhaltinger, P. Twaróg, and M. Bogdan. Modified versions of bayesian information criterion for genome-wide association studies. *Computational Statistics & Data Analysis*, 56(5): 1038–1051, 2012.

B. J. Gu, P. N. Baird, K. A. Vessey, K. K. Skarratt, E. L. Fletcher, S. J. Fuller, A. J. Richardson, R. H. Guymer, and J. S. Wiley. A rare functional haplotype of the p2rx4 and p2rx7 genes leads to loss of innate phagocytosis and confers increased risk of age-related macular degeneration. *The FASEB Journal*, 27(4):1479–1487, 2013.

F. Han and W. Pan. A composite likelihood approach to latent multivariate gaussian modeling of snp data with application to genetic association testing. *Biometrics*, 68(1):307–315, 2012.

P. R. Hansen, A. Lunde, and J. M. Nason. The model confidence set. *Econometrica*, 79(2):453–497, 2011.

J. B. Jonas. Global prevalence of age-related macular degeneration. *Lancet Global Health*, 2(2): e65–e66, 2014.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. springer, 2006.

J. Li, W. Zhong, R. Li, and R. Wu. A fast algorithm for detecting gene–gene interactions in genome-wide association studies. *The Annals of Applied Statistics*, 8(4):2292–2318, 2014.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

W. Pan. Asymptotic tests of association with multiple snps in linkage disequilibrium. *Genetic Epidemiology*, 33(6):497–507, 2009.

M. Y. Park and T. Hastie. Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30–50, 2008.

G. Qian and C. Field. Using mcmc for logistic regression model selection involving large number of candidate models. In *Monte Carlo and Quasi-Monte Carlo Methods 2000*, pages 460–474. Springer, 2002.

X. Wen. Bayesian model comparison in genetic association analysis: linear mixed modeling and snp set testing. *Biostatistics*, 16:701–712, 2015.

T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.