MODEL SELECTION CONFIDENCE SETS
BY LIKELIHOOD RATIO TESTING

Chao Zheng, Davide Ferrari and Yuhong Yang

*Lancaster University, Free University of Bozen-Bolzano
and University of Minnesota*

Abstract: The traditional activity of model selection aims at discovering a single
model superior to other candidate models. In the presence of pronounced noise,
however, multiple models are often found to explain the same data equally well. To
resolve this model selection ambiguity, we introduce the general approach of model
selection confidence sets (MSCSs) based on likelihood ratio testing. A MSCS is
defined as a list of models statistically indistinguishable from the true model at a
user-specified level of confidence, which extends the familiar notion of confidence in-
tervals to the model-selection framework. Our approach guarantees asymptotically
correct coverage probability of the true model when both sample size and model
dimension increase. We derive conditions under which the MSCS contains all the
relevant information about the true model structure. In addition, we propose natu-
ral statistics based on the MSCS to measure importance of variables in a principled
way that accounts for the overall model uncertainty. When the space of feasible
models is large, MSCS is implemented by an adaptive stochastic search algorithm
which samples MSCS models with high probability. The MSCS methodology is
illustrated through numerical experiments on synthetic and real data examples.

Key words and phrases: Adaptive sampling, likelihood ratio test, model selection
confidence set, optimal detectability condition.

1. Introduction

Likelihood inference is a centerpiece of statistical theory and plays an im-
portant role in many research fields. Numerous methods relying on likelihood
objective functions have been developed in the literature of model selection, rang-
ing from classic information criteria to more recent sparsity-inducing penaliza-
tion methods; see McQuarrie and Tsai (1998), Claeskens and Hjort (2008) and
Bühlmann and van de Geer (2011) for book-length expositions. In the presence
of noise in data, however, it is typically difficult to declare a single model sig-
nificantly superior to all possible competitors due to the prevailing effect of the
model selection uncertainty. In this situation, multiple or even a large number

of models may be equally supported by the data, so that any selection proce-
dure is likely to pick a single model from a large set of more or less equivalent
models. Clearly, this implies the tossing away valuable information; for example,
in regression analysis, alternative combinations of predictors may be discarded,
whilst such combinations may contain scientifically valid explanations of the phe-
nomenon under examination.

Motivated by these issues, there has been a growing interest in developing
statistical measures of model selection uncertainty. We propose to construct
a model selection confidence set (MSCS), defined as a set of models indistin-
guishable from the true model at a user-defined confidence level. Simply put,
the MSCS extends the familiar frequentist notion of confidence intervals to the
model-selection framework. Ferrari and Yang (2015) first introduced confidence
sets for variable selection in the context of linear models by F-testing. They
achieve the exact coverage probability for the globally optimal model from the
model space. This is the first work introducing confidence sets in the frequantist
sense for variable selection. In their framework the number of predictors can
grow with the sample size, so that the number of potentially useful models is
allowed to be large.

Different from the MSCS of Ferrari and Yang (2015), Hansen, Lunde and
Nason (2011) studied the methodology of model confidence sets. Their approach
builds on classic step-down procedures for multiple hypothesis testing (Lehmann
and Romano (2005); Romano and Wold (2005)). Starting from a fixed user-
defined set of models, which has limited sizes, they carry out step-wise equivalence
testing under a user-defined loss function, followed by an elimination rule to
drop the worst performing models. Previously, Shimodaira (1998) constructed
confidence sets containing models with AIC values near the smallest among the
candidate models. We refer to the Section 6 of Ferrari and Yang (2015) for a
detailed discussion.

In this paper, we introduce a general methodology to construct model selec-
tion confidence sets via likelihood inference. We begin by considering a full model
with p variables to form a reference model or full model, where p is required to be
less than n. This preliminary step can be achieved by any over-consistent model
screening method that selects the relevant variables plus a few other variables.
We then test candidate sub-models against the full model by a likelihood ratio
test (LRT) at the significance level $0 < \alpha < 1$. The MSCS is formed by all
the candidate models that survive the LRT screening. This way of construction
guarantees that the globally optimal model is included in the MSCS with prob-

ability at least $1 - \alpha$ as the sample size increases (under appropriate regularity conditions). From a theoretical viewpoint, we investigate the condition for the MSCS to contain all the relevant information about the model structure when both p and n diverge. Since in practice the MSCS cannot be computed by exhaustive search unless p is very small, when the model space is moderate or large, we propose a stochastic algorithm (MSCS-AS) that samples MSCS models with high probability.

The proposed MSCS methodology can be used for various tasks in support of the model selection activity. First, given a model selected from some rule, one can immediately use it to check if such a model is too parsimonious in terms of missing important variables. Second, the frequency of variables in the MSCS can be used to rank their usefulness in a principled way that accounts for the model selection uncertainty. Third, the MSCS and the associated importance measures can be used to narrow down the list of candidate models by considering the most important variables.

The rest of the paper is organized as follows. In Section 2, we describe the main MSCS methodology and study the conditions needed to learn the true underlying model structure. In the same section, we propose a measure of importance for the individual variables. In Section 3, we give an adaptive sampling algorithm that implements the MSCS methodology. In Section 4, we study the finite sample properties of MSCS by Monte Carlo simulations for various models. In Section 5, we illustrate the MSCS procedure using the European E.coli outbreak data and the Australian breast cancer family study data. In Section 6, we conclude and give final remarks. Technical proofs are deferred to the Appendix.

2. Model Selection Confidence Sets

Consider independent observations, $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, from a family of models indexed by the parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T \in \boldsymbol{\Theta}$, with corresponding log-likelihood function $\ell_n(\boldsymbol{y}, \boldsymbol{\theta})$. Each parameter element $\theta_j \in \boldsymbol{\theta}$ describes a possibly relevant part of the overall model structure. We suppose that only a subset of $\boldsymbol{\theta}$ is useful for describing the data, while the others are regarded as unnecessary. A generic model index $\boldsymbol{\gamma}$ is defined as a subset of indexes in $\{1, \ldots, p\}$ and we write the corresponding parameter space as $\boldsymbol{\Theta}_{\boldsymbol{\gamma}}$. Let $\boldsymbol{\theta}_{\boldsymbol{\gamma}} \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}}$ be a parameter with the model $\boldsymbol{\gamma}$, and let $p_{\boldsymbol{\gamma}} = \mathrm{card}(\boldsymbol{\gamma})$ denote the number of elements in $\boldsymbol{\gamma}$. The true parameter vector and the true model is denoted by $\boldsymbol{\theta}^*$ and $\boldsymbol{\gamma}^*$, repectively, while the full model with p parameters is denoted by $\boldsymbol{\gamma}_f$. The space of feasible candi-

date models is Γ, which contains the true model γ^*. The cardinality of Γ may be as large as 2^p; however it may also be restricted in some special problems.

In the rest of the paper, we assume $p < n$, but p is allowed to slowly grow with n, reflecting the notion that with more observations available, the statistician is tempted to introduce additional variables into the model. For simplicity, we omit the sub-index n when it is clear from the context. In what follows, we use "$\lesssim$" to denote that the left hand side is bounded by the right hand side up to some positive constant independent of n. We write $a \gtrsim b$ if $b \lesssim a$.

2.1. Construction by likelihood ratio testing

A MSCS is constructed from the known models space, Γ, and a criterion to assess models in Γ empirically. To screen out implausible models in the context of maximum likelihood estimation, it is natural to use the likelihood ratio test. Given a candidate model γ, we consider testing the null hypothesis $H_0 : \boldsymbol{\theta}^* \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}}$ against the alternative hypothesis $H_1 : \boldsymbol{\theta}^* \notin \boldsymbol{\Theta}_{\boldsymbol{\gamma}}$. Then model γ is rejected if

$$\Lambda_{\boldsymbol{\gamma}} \equiv 2 \left\{ \ell_n(\boldsymbol{\theta}_{\boldsymbol{\gamma}_f}) - \ell_n(\boldsymbol{\theta}_{\boldsymbol{\gamma}}) \right\} \geq q(\alpha; p - p_{\boldsymbol{\gamma}}), \tag{2.1}$$

where: $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}_f}$ denote, respectively, the MLEs for the candidate and full models; $\ell_n(\cdot)$ is the log-likelihood function; $q(\alpha; d)$ is the upper α-quantile for the central chi-squared distribution with d degree of freedom. The $(1 - \alpha)100\%$–MSCS is defined by the set of all models surviving the LRT screening:

$$\widehat{\Gamma}_\alpha \equiv \left\{ \boldsymbol{\gamma} \in \Gamma : \Lambda_{\boldsymbol{\gamma}} \leq q(\alpha; p - p_{\boldsymbol{\gamma}}) \right\}. \tag{2.2}$$

The LRT procedure is applied to all models $\boldsymbol{\gamma} \in \Gamma$: If a model is rejected, then we have evidence that it is too parsimonious in the sense that it is likely to miss at least one important variable. By default, the full model $\boldsymbol{\gamma}_f$ is included in $\widehat{\Gamma}_\alpha$.

When p is fixed, and γ^* is a proper subset of γ, the limiting null distribution of the LRT statistics $\Lambda_{\boldsymbol{\gamma}}$ is a central chi-square distribution, as follows from Wilks theorem (e.g., see van der Vaart (2000)). By construction, this implies that the true model is in the MSCS with probability approximately $1 - \alpha$ in large samples. Specifically, if the true model is not the full model ($\gamma^* \neq \gamma_f$), we have:

$$\lim_{n \to \infty} P(\gamma^* \in \widehat{\Gamma}_\alpha) = 1 - \alpha. \tag{2.3}$$

If $\gamma^* = \gamma_f$, then $P(\gamma^* \in \widehat{\Gamma}_\alpha) = 1$.

When p increases with n, similar Wilks-type results are given by Portnoy (1988), Murphy (1993), Fan and Peng (2004) and Fan and Zhou (2015) for exponential family models, Cox regression, penalized likelihood and goodness of spurious fit for GLMs. These results yield asymptotic coverage probability as in

(2.3). Spokoiny (2012, 2013) establish Wilks-type behaviours for rather general families of models, which quantify and explicitly describe the error term in the approximation of the likelihood ratio statistics under mild regularity conditions on the parametric family.

We remark that although the MSCS includes $\boldsymbol{\gamma}^*$ with at least probability $1 - \alpha$, one cannot simply conclude that a variable is important just because it appears in some of the models in $\widehat{\Gamma}_\alpha$. Actually, unimportant variables tend to appear with a respectable frequency in the MSCS models since larger models containing the true model plus other irrelevant variables are likely to survive the LRT screening.

2.2. Asymptotic detectability

In this section, we study the conditions under which the variables in the true model appear with large frequency in MSCS. The results presented in this section extend the analysis given by Ferrari and Yang (2015) for linear models.

Definition 1 (Asymptotic detectability). *The MSCS $(\widehat{\Gamma}_\alpha)$ is said to asymptotically detect all the true variables, if all the variables in the true model $\boldsymbol{\gamma}^*$ are included in each of the models in $\widehat{\Gamma}_\alpha$, with probability going to* 1.

The concept of detectability is closely related to the power of the LRT. In the fixed p scenario when a candidate model $\boldsymbol{\gamma}$ misses at least one important variable, under appropriate regularity conditions ensuring asymptotic normality of the MLE, the $\Lambda_{\boldsymbol{\gamma}}$ converges in distribution to a non-central chi-square random variable with degree of freedom $d_{\boldsymbol{\gamma}} = p - p_{\boldsymbol{\gamma}}$. Let $\boldsymbol{\theta}_{\boldsymbol{\gamma}}^*$ denote the parameter value in the model γ that minimizes the Kullback-Leibler divergence from the true density. Then the non-centrality parameter of the asymptotic chi-square distribution is $\delta_{\boldsymbol{\gamma}} = (\boldsymbol{\theta}_{\boldsymbol{\gamma}}^* - \boldsymbol{\theta}^*)^T \mathfrak{F}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}_{\boldsymbol{\gamma}}^* - \boldsymbol{\theta}^*)$, where $\mathfrak{F}(\boldsymbol{\theta}) = -E\left\{\partial^2 \ell_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^2\right\}$ is the Fisher information matrix. In the normal regression case, as in Ferrari and Yang (2015), the asymptotic distribution is exact. Clearly, when applying LRT to a model $\boldsymbol{\gamma}$, a large value of $\delta_{\boldsymbol{\gamma}}$ makes it easier to reject $\boldsymbol{\gamma}$.

In the following theorem, we show that the limiting non-central chi-square alternative distribution is still valid under certain conditions for an exponential model, where $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are i.i.d. observations from the pdf

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}^T \boldsymbol{y} - A(\boldsymbol{\theta})\right). \tag{2.4}$$

with respect to a sigma-finite dominating measure. For this model, the true model $\boldsymbol{\gamma}^*$ is defined as the indexes of non-zero component of $\boldsymbol{\theta}$. Appropriate generalizations of this result may be derived for other models, but are not pursued

there.

Theorem 1. *Assume conditions* $(A1) - (A3)$ *given in the Appendix. Let* $\boldsymbol{\gamma}$ *be a model missing at least one variable in the true model* $\boldsymbol{\gamma}^*$ *and* $d_{\boldsymbol{\gamma}} \to \infty$ *as* n *grows. If* $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_{\boldsymbol{\gamma}}^*\| \gtrsim \sqrt{p/n}$ *and* $p = o(n^{2/3})$, *then for model (2.4) we have*

$$\frac{\Lambda_{\boldsymbol{\gamma}} - (d_{\boldsymbol{\gamma}} + \delta_{\boldsymbol{\gamma}})}{\sqrt{2d_{\boldsymbol{\gamma}} + 4\delta_{\boldsymbol{\gamma}}}} \xrightarrow{\mathscr{D}} \mathbb{N}_1(0, 1), \quad as \ n \to \infty. \tag{2.5}$$

Write $X_{\delta,d}$ as a chi-square random variable with degree of freedom d and non-centrality parameter δ. As $(X_{\delta,d} - d - \delta)/\sqrt{2(d + 2\delta)}$ converges to a standard normal when $d \to \infty$, in view of (2.5), $\Lambda_{\boldsymbol{\gamma}}$ is approximately non-central chi-square variable with degree of freedom $d_{\boldsymbol{\gamma}}$ and non-centrality parameter $\delta_{\boldsymbol{\gamma}}$.

The non-centrality parameter $\delta_{\boldsymbol{\gamma}}$ can be interpreted a discrepancy measure due to missing important variables in the true model. From this viewpoint, the relative magnitude of $\delta_{\boldsymbol{\gamma}}$ provides us with some insight on how informative is the data in relation to the feasibility of the model selection task. Let $K_n(s) = s \log(p/s)$. With $p \to \infty$, it is typically the case that the true model dimension is bounded away significantly from p. We assume that $d_{\boldsymbol{\gamma}}$ increases to ∞ (however slowly) uniformly for the candidate models. We have sufficient conditions involving $\delta_{\boldsymbol{\gamma}}$ for detectability in the general parametric setting, which includes model (2.4).

Theorem 2. *Let* Γ_u *denote the set of models missing at least one of the true variables. If it holds that for all* $\boldsymbol{\gamma} \in \Gamma_u$, *we have*

$$|P\left(\Lambda_{\boldsymbol{\gamma}} \leq q(\alpha; k)\right) - P\left(X_{\delta_{\boldsymbol{\gamma}}, k} \leq q(\alpha; k)\right)| \leq c_1 \exp\left(-c_2 K_n(k)\right), \tag{2.6}$$

where c_1 *and* c_2 *are positive constants. A sufficient condition for asymptotic detectability is*

$$\min_{\boldsymbol{\gamma} \in \Gamma_u} \frac{\delta_{\boldsymbol{\gamma}}}{K_n(d_{\boldsymbol{\gamma}})} > B, \tag{2.7}$$

for some large enough positive constant B.

This additional assumption requires an exponential probability bound for the chi-square approximation of LRT statistics with model misspecification. We refer to Theorem 3.10 and Proposition B.1 in Spokoiny (2013) for the chi-square approximation, where a similar bound can be achieved for certain i.i.d and regression models. Moreover, consider normal linear regression as in Ferrari and Yang (2015), where $\left(RSS_{\boldsymbol{\gamma}} - RSS_{\boldsymbol{\gamma}_f}\right)/d_{\boldsymbol{\gamma}}$, the numerator of their F-test, follows an exactly non-central chi-square distribution and the assumption is trivially satisfied.

2.3. Sharpness of the sufficient condition for detectability

Theorem 2 is an extension of Theorem 2.3 in Ferrari and Yang (2015) where, in the context of normal linear regression, a sufficient condition for detectbility is given as $\min_{\gamma \in \Gamma_u} \delta_\gamma / \left\{ \xi_n + \sqrt{K_n(d_\gamma)} \right\}$ is greater than some large enough constant, where $\xi_n \to \infty$ is any arbitrarily slowly growing sequence. It turns out their condition is in fact not sufficient, an error occurred in its derivation. A sufficient condition is that $\min_{\gamma \in \Gamma_u} \delta_\gamma / K_n(d_\gamma)$ is larger than some constant as in (2.7). In this subsection, we show this sufficient condition cannot be generally improved. We focus on the normal regression case here. A generalization to other models, e.g. GLMs, can be done similarly with additional technical developments.

The detectability condition relates to the size of the coefficients. For the following results, we assume the sparse Riesz condition (SRC) as in Zhang (2010) holds and consider that $0 < \alpha < 1/2$.

Let $r^* \leq p/2$ be a positive integer, the number of non-zero coefficients in the true model. Write $f_{\boldsymbol{\beta}}(\mathbf{X}) = \mathbf{X}^T \boldsymbol{\beta} = \sum_{j=1}^p \beta_j X_j$ and let $\mathcal{B} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_0 = r^*$ and $\|f_{\boldsymbol{\beta}}\|_n^2 \leq c K_n(r^*)\}$ for some small constant $c > 0$, where $\|\cdot\|_0$ denotes the ℓ_0-norm, and $\|f_{\boldsymbol{\beta}}\|_n^2 = \sum_{i=1}^n f_{\boldsymbol{\beta}}^2(\mathbf{X}_i)$ with $\mathbf{X}_i$, the covariate vector for the i-th observation. This represents all linear regression models $\boldsymbol{\gamma}$ with only r^* non-zero coefficients.

Theorem 3. *Let $\mathcal{D}$ denote the event that all the variables in $\boldsymbol{\gamma}*$ are included in each of the models in the MSCS. Then when c is small enough, we have*

$$\limsup_{n \to \infty} \inf_{\boldsymbol{\beta} \in \mathcal{B}} P_{\boldsymbol{\beta}}(\mathcal{D}) < 1.$$

From the theorem, for the true models of dimension r^* with $\|f_{\boldsymbol{\beta}}\|_n^2 \leq c K_n(r^*)$, detection of the true terms is impossible in a proper minimax sense. For instance, for the model γ that contains all the wrong variables and none of the true variables, it results in $d_\gamma = r^*$ and the noncentrality parameter is of order $K_n(d_{\boldsymbol{\gamma}})$. This matches the lower bound requirement (2.7) in Theorem 2. The sufficient condition (2.7) for detectability cannot be generally weakened.

2.4. Inclusion importance

Under the detectability conditions in Theorem 2, the MSCS includes all the relevant information concerning the model selection variability. Thus, a natural measure for ranking the importance of each parameter element $\theta_k \in \boldsymbol{\theta}$ is its relative frequency over all the MSCS models. This suggests the following.

Definition 2 (Inclusion Importance)**.** *The inclusion importance (II) for any*

$\theta_k \in \boldsymbol{\theta}$ *is*

$$II_k = \sum_{\gamma \in \widehat{\Gamma}_\alpha} \frac{I(k \in \boldsymbol{\gamma})}{\operatorname{card}(\widehat{\Gamma}_\alpha)}. \tag{2.8}$$

When θ_k appears in all MSCS models, its importance is $II_k = 1$, θ_k is most likely part of the true model.

As mentioned, a variable cannot be declared relevant just because it has a non-zero importance index, unimportant variables of $\boldsymbol{\theta}$ tend to appear in the MSCS with frequency near $1/2$. The reason is that when a small model is included in $\widehat{\Gamma}_\alpha$, larger models containing the same variables plus some others tend to be included via the LRT by construction.

Theorem 4. *If the conditions in Theorem* 2 *are satisfied, we have:*

(i) $\lim_{n \to \infty} P(II_k = 1) = 1$, *for all* $k \in \boldsymbol{\gamma}^*$;

(ii) $\lim_{n \to \infty} P\left(II_k > (1/2) + \Delta\right) \leq \alpha(1 + 2\Delta)/4\Delta$, *for all* $k \notin \boldsymbol{\gamma}^*$, *where* $0 < \Delta < 1/2$.

With sufficient information to learn all the the relevant variables of the true model, we expect that their importance to be close to 1, while the unimportant variables are not likely exceeding by much the value 0.5. The upper bound in Theorem 4 can be used as a guidance to control the error probability of over selection. One could set the error probability $\epsilon = \alpha(1+2\Delta)/(4\Delta)$ to be some small number and then find the corresponding Δ so as to use II for an understanding if a variable is important. For example, if the significance level is $\alpha = 0.05$, setting $\Delta = 1/6$ implies $\epsilon \leq 0.1$.

2.5. The multivariate normal location model

In this section we consider the special case of the multivariate normal distribution with unknown location. Let $\mathbf{Y}$ be p-variate normal $\mathcal{N}_p(\boldsymbol{\mu}, I_p)$. The pdf with form (2.4) can be obtained by setting $\boldsymbol{\theta} = \boldsymbol{\mu}$ and the cumulant generating function is $A(\boldsymbol{\theta}) = \boldsymbol{\theta}^T \boldsymbol{\theta}/2 + p \log(2\pi)/2$. Assume the true parameter $\boldsymbol{\theta}^*$ is sparse with $p_{\boldsymbol{\gamma}^*} = o(p)$. The model space Γ then has cardinality 2^p.

For a misspecified model $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_u$, the corresponding non-centrality parameter is $\delta_{\boldsymbol{\gamma}} = n \sum_{j \in \boldsymbol{\gamma}^* \text{ and } j \notin \gamma} \boldsymbol{\theta}_j^{*2}$. A large value of $\delta_{\boldsymbol{\gamma}}$ enables us to detect inadequacy of such models. For example, the asymptotic detectability conditions in (2.7) states that as long as the minimum signal is large enough, $\min(|\theta_i|, i \in \boldsymbol{\gamma}^*) > B\sqrt{p/n}$ for a large enough positive constant B, then all the models in MSCS are expected to contain all the nonzero parameters with probability going to 1 as $n \to \infty$.

Otherwise if the size of some non-zero parameters in $\boldsymbol{\theta}^*$ is too small, the LRT has not enough power to screen out some wrong models.

3. Implementation by Adaptive Sampling

Testing all the models in Γ is computationally challenging unless p is small, since the cardinality of the model space can grow exponentially in p. Thus, in order to find models in the MSCS it seems natural to turn to sampling methods. Let $\mathbf{U} = (u_1, \ldots, u_p)^T \sim p(\mathbf{u}; \boldsymbol{\omega})$ be a random binary vector representing a model sampled from Γ ($u_j = 1$ if the jth variable is included in the model), and $p(\cdot; \boldsymbol{\omega})$ is a user-defined pmf indexed by $\boldsymbol{\omega}$. Our main objective is to choose a value of the parameter $\boldsymbol{\omega}$ that maximizes the probability to sample MSCS models,

$$P(\mathbf{U} \in \widehat{\Gamma}_\alpha) = \sum_{\mathbf{u} \in \Gamma} p(\mathbf{u}; \boldsymbol{\omega}) I(\mathbf{u} \in \widehat{\Gamma}_\alpha). \tag{3.1}$$

Note one is unlikely to find models in $\widehat{\Gamma}_\alpha$ just by sampling from some arbitrary pmf $p(\cdot; \boldsymbol{\omega})$, unless α is sufficiently small. Thus, given a target significance level $\alpha = \alpha^*$ (e.g. 0.05) we propose to start from some small initial confidence level, say $\alpha^{(0)}$, and then construct a sequence of significance levels, $0 < \alpha^{(0)} \le \alpha^{(1)} \le \cdots \le \alpha^*$, corresponding to sampling distributions $p(\cdot; \boldsymbol{\omega}^{(0)}), p(\cdot; \boldsymbol{\omega}^{(1)}), \ldots, p(\cdot; \boldsymbol{\omega}^{(*)})$ increasingly concentrated on the target subspace $\widehat{\Gamma}_{\alpha^*}$.

At each step $t \ge 0$ of our algorithm, the parameter $\boldsymbol{\omega}$ of the sampling distribution is retrieved by a weighted likelihood approach. We generate B models $\{\mathbf{u}_b^{(t)}, b = 1, \ldots, B\}$ from $p(\cdot; \hat{\boldsymbol{\omega}}^{(t-1)})$ and then compute

$$\hat{\boldsymbol{\omega}}^{(t)} = \underset{\boldsymbol{\omega}}{\operatorname{argmax}} \sum_{b=1}^{B} I(\mathbf{u}_b^{(t)} \in \widehat{\Gamma}_{\alpha^{(t-1)}}) p(\mathbf{u}_b^{(t)}; \boldsymbol{\omega}). \tag{3.2}$$

This finds the pmf $p(\cdot; \hat{\boldsymbol{\omega}}^{(t)})$ closest to the best subset of previously sampled models in terms of their resemblance to MSCS models. As t increases and $\alpha^{(t)}$ gradually gets closer to α^*, $p(\cdot; \hat{\boldsymbol{\omega}}^{(t)})$ tends to assign larger probability to models in $\widehat{\Gamma}_\alpha$.

Since this procedure is useful only when the indicator $I(\mathbf{u}_b^{(t-1)} \in \widehat{\Gamma}_{\alpha^{(t-1)}}) = 1$ for a sufficiently large fraction of sampled models, $\mathbf{u}_1^{(t-1)}, \ldots, \mathbf{u}_B^{(t-1)}$, we propose to increase adaptively the significance level as $\alpha^{(t)} = \min\{\text{p-val}_{\lfloor (1-\zeta)B \rfloor}^{(t)}, \alpha^*\}$, $0 < \zeta < 1$, where $\text{p-val}_{\lfloor (1-\zeta)B \rfloor}$ is the empirical $(1-\zeta)$-quantile computed from the distribution of p-values. This ensures that the event $\{\boldsymbol{\gamma}_b^{(t)} \in \widehat{\Gamma}_{\alpha^{(t)}}\}$ is not too rare and occurs with probability of approximately ζ. This approach is closely related to cross-entropy (CE) sampling. See Rubinstein and Kroese (2004) for a book-

length exposition on this topic and Costa, Jones and Kroese (2007) for convergence analysis. In our implementation, we use $p(\mathbf{u}; \boldsymbol{\omega}) = \prod_{j=1}^{p} \omega_j^{u_j} (1 - \omega_j)^{1-u_j}$, which gives a closed-form solution to (3.2) and leads to a fast algorithm; our numerical experiments showed reliable results with relatively fast convergence. Other choices for $p(\mathbf{u}; \boldsymbol{\omega})$ may enhance the performance of the algorithm but are not pursued here.

Algorithm 1 :   MSCS construction by adaptive sampling (MSCS-AS)

---

0. Initialize $t = 0$ (iteration counter) and $\hat{\boldsymbol{\omega}}^{(0)}$ (parameter vector for pmf $p(\mathbf{u}, \boldsymbol{\omega})$).

1. Set $t \leftarrow t + 1$. Generate $S^{(t)} = \{\mathbf{u}_1^{(t)}, \ldots, \mathbf{u}_B^{(t)}\}$ from $p(\cdot; \hat{\boldsymbol{\omega}}^{(t-1)})$, and compute the sorted p-values, p-val$_{(1)}^{(t)} \leq \cdots \leq$ p-val$_{(B)}^{(t)}$, by the LRT defined in (2.1).

2. Update $\alpha^{(t)} = \min\left(\text{p-val}_{\lfloor(1-\varsigma)B\rfloor}^{(t)}, \alpha^*\right)$.

3. Use models sampled in Step 1, maximize the weighted likelihood as in(3.2) by computing:

$$c_j^{(t)} = \frac{\sum_{b=1}^{B} I\{\text{p-val}_{(b)}^{(t)} > \alpha^{(t)}, \hat{\theta}_j \in \mathbf{u}_b^{(t)}\}}{\sum_{b=1}^{B} I\{\text{p-val}_{(b)}^{(t)} > \alpha^{(t)}\}}, \quad j = 1, \ldots, p,$$

where $\{\theta_j \in \mathbf{u}\}$ denotes the event that the variable θ_j appears in model $\mathbf{u}$.

4. Update $\hat{\omega}_j^{(t)} \leftarrow \xi c_j^{(t)} + (1 - \xi)\hat{\omega}_j^{(t-1)}$ for some constant $0 < \xi < 1$.

5. Repeat Steps 1–4 until $\alpha^{(t-d)} = \cdots = \alpha^{(t)} = \alpha^*$, for some d (e.g. $d = 10$). The final MSCS is obtained by drawing $B^{(T)}$ models from $p(\mathbf{u}; \hat{\boldsymbol{\omega}}^{(T)})$, where T denotes the last iteration.

Here, Step 4 carries out smoothing at each iteration; if $\xi = 1$, the algorithm avoids smoothing. In our simulations, we found that $\xi < 1$ performs better than the non-smooth update with $\xi = 1$ since it prevents occurrences of too many zeros and ones in situations where p is moderate or large. Smoothing avoids local optima where some model variables do not have the chance to be selected, while others are always selected. In our experience, the MSCS-AS algorithm is robust to the choice of ξ, with $\xi = 0.2$ performing well across all our numerical examples.

The MSCS-AS algorithm requires setting the initial weights $\hat{\boldsymbol{\omega}}^{(0)}$, and the number of models sampled at each iteration B. We found that the procedure is quite robust to the choices of such parameters. When no prior information on

inclusion importance is available, the initial probabilities $\hat{\boldsymbol{\omega}}^{(0)}$ can be set as $\hat{\boldsymbol{\omega}}^{(0)} = (0.5, \ldots, 0.5)$. The performance of the method, however, can be improved by assigning larger weights to variables that are known to contain more information about the true model. The number of models B generated in each iteration should be decided based on affordable computational resources. However, if B is too small, this will affect the accuracy of the weighted likelihood criterion (3.2). In our numerical examples we set $B = 300$.

The constant ζ prevents overly small p-values in the first few iterations; thus, it ensures a balanced growth of $\alpha^{(t)}$ and guides the sampling process towards the MSCS models. The parameter ζ governs the trade-off between exploration and exploitation of the model space Γ and it should be also fixed based on the available computational resources. In our simulations, $\zeta = 0.25$ was found to work well and is compatible with choices of analogous parameters often found in the CE literature.

4. Monte Carlo Experiments

4.1. MSCS construction by exhaustive search

In this subsection, we study the finite sample properties of MSCSs constructed by exhaustive search on the model space Γ. We generated samples from four models:

Model 1: p-variate normal with unknown location, $\mathbf{Y} = (Y_1, \ldots, Y_p)^T \sim \mathcal{N}_p(\boldsymbol{\theta}, I)$.

Model 2: p-variate normal $\mathbf{Y} = (Y_1, \ldots, Y_p)^T \sim \mathcal{N}_p(0, \Sigma)$, with unknown covariance matrix Σ. Additionally, $(Y_1, \ldots, Y_p)$ can be partitioned into independent subvectors ($\Sigma_{i,j} = 0$ if Y_i and Y_j belong to different subvectors).

Model 3: Logistic regression: $Y|\mathbf{x} \sim \text{Bernoulli}(\pi(\mathbf{x}))$, $\text{Logit}(\pi(x)) = -\mathbf{x}^T\boldsymbol{\theta}$, $\boldsymbol{\theta} \in \mathbb{R}^p$.

Model 4: Poisson regression: $Y|\mathbf{x} \sim \text{Poisson}(\lambda(\mathbf{x}))$, $\log(\lambda(\mathbf{x})) = -\mathbf{x}^T\boldsymbol{\theta}$, $\boldsymbol{\theta} \in \mathbb{R}^p$.

In Models 3 and 4, the vectors of covariates were sampled from multivariate normal distribution $\mathcal{N}_p(0, I)$ at each Monte Carlo run. For all the models we considered two scenarios.

Setting 1: (constant parameter size): The first $p/2$ parameters have the same size and the others are equal to 0. Specifically, $\theta_j = \psi$, $j =$

$1, \ldots, p/2$, and $\theta_j = 0$, $j = p/2 + 1, \ldots, p$, where the parameters size ψ is set to be 1 for Models 1 and 3 and 0.2 for Models 4, respectively. For Model 2, $(Y_1, \ldots, Y_p)$ is partitioned into $p/2 + 1$ groups as $\{Y_1, \ldots, Y_{p/2}\}, \{Y_{p/2+1}\}, \ldots, \{Y_p\}$. We set $\Sigma_{i,j} = 1$ if $i = j$; $\Sigma_{i,j} = 0.5$, $1 \le i < j \le p/2$, and $\Sigma_{i,j} = 0$ otherwise.

Setting 2: (decreasing parameter size): The first $p/2$ coefficients have decreasing size and the others are equal to 0. Specifically, $\theta_j = \psi/j$, $j = 1, \ldots, p/2$, and $\theta_j = 0$, $j = p/2 + 1, \ldots, p$, where ψ is set as $1, 2$ and 0.4 for Models 1, 3 and 4, respectively. For Model 2, $(Y_1, \ldots, Y_p)$ is partitioned into $p/2 + 1$ groups as $\{Y_1, \ldots, Y_{p/2}\}$, $\{Y_{p/2+1}\}, \ldots, \{Y_p\}$. We set $\Sigma_{i,j} = 1$ if $i = j$; $\Sigma_{i,j} = 0.5/|i - j|$ if $1 \le i < j \le p/2$, and $\Sigma_{i,j} = 0$ otherwise.

These above settings are designed to achieve a small signal relative to the noise, so that the resulting data are affected by model-selection uncertainty. The model space consists of 2^p models for Models 1, 3 and 4. For Model 2, the model space is equivalent to all the possible partitions of the set $\{1, 2, \ldots, p\}$ and its cardinality can be given by the Bell number B_p. In the following simulations, we used $p = 6, 8$ corresponding to $B_6 = 203$ and $B_8 = 4{,}140$.

Tables 1 and 2 show Monte Carlo estimates for the coverage probability and cardinality of MSCS corresponding to different sample sizes, n, number of predictors, p, at the 90, 95 and 99% confidence levels. As one expects, the cardinality of the MSCS grows as α decreases, while it increases rapidly with p, especially when the sample size n is relatively small. This reflects the situation where the data contain too much noise and the subsequent model selection variability is pronounced. The cardinality of the MSCS drops quickly as n increases.

In most cases, the true coverage probability is quite close to the nominal confidence level. And will be improved in general as the sample szies increases. The true coverage probability tends to be more off from the nominal level when the size of the true parameters is decreasing (Setting 2). Clearly, in such settings model selection is more challenging, which leads to a increased cardinality of the MSCS but maintains the same coverage probability.

4.2. MSCS construction by stochastic search

In this subsection, we study the performance of the MSCS-AS algorithm described in Section 3. We generated data from Models 3 and 4 (Poisson and Logistic regression models) using a setting similar to that in Fan and Lv (2011). For

Table 1. Monte Carlo estimates of MSCS coverage probability and cardinality by exhaustive search for Model 1 and Model 2 under varying confidence level $1 - \alpha$, sample size n, and number of variables p. Results based on 500 MC runs.

			Setting 1				Setting 2			
		$n =$	100		250		100		250	
	α					Model 1				
		$p =$	8	12	8	12	8	12	8	12
Coverage (%)	0.10		91.2	90.0	89.8	90.2	91.2	89.4	91.0	88.6
	0.05		94.4	95.8	94.4	96.2	94.8	95.4	95.0	94.4
	0.01		98.4	99.8	99.2	98.4	99.4	98.8	99.0	99.4
Cardinality	0.10		14.5	58.1	14.4	58.3	22.7	241.3	15.5	101.3
	0.05		15.3	61.0	15.2	61.2	27.2	319.0	16.8	126.5
	0.01		15.8	63.5	15.9	63.4	37.1	497.8	19.4	184.6
						Model 2				
		$p =$	6	8	6	8	6	8	6	8
Coverage (%)	0.10		89.2	88.4	89.2	91.0	89.0	87.6	88.4	88.8
	0.05		94.6	94.8	94.2	94.4	94.4	93.2	94.4	94.4
	0.01		98.4	97.8	98.8	99.4	98.4	99.0	98.6	99.2
Cardinality	0.10		13.5	47.1	13.5	47.0	13.8	57.1	13.4	46.4
	0.05		14.3	50.7	14.2	47.3	15.1	72.7	14.2	49.2
	0.01		15.6	57.9	14.8	46.9	18.8	126.6	14.8	51.5

both models, we set $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^T, 0, 0, \ldots, 0\right)^T$, with $\boldsymbol{\theta}_1 = (2.5, -1.9, 2.8, -2.2, 3)^T$ in Model 3 and $\boldsymbol{\theta}_1 = (1.25, -0.95, 0.9, -1.1, 0.6)^T$ in Model 4. The vector of covariates were sampled from a multivariate normal $\mathcal{N}_p(0, \Sigma)$ at each Monte Carlo run, where Σ had elements $\Sigma_{i,j} = 0.5^{|i-j|}$, $i, j = 1, 2, \ldots, p$. To illustrate the capability to handle large p, we show the results for (n, p) equal to $(200, 100)$ and $(1,000, 500)$. The initial weights for the MSCS-AS algorithm were $\hat{\boldsymbol{\omega}}^{(0)} = (0.5, \ldots, 0.5)$, corresponding to lack of prior information about predictors' importance. The remaining tuning parameters were set as $\zeta = 0.25$, $\alpha^* = 0.05$, and $\xi = 0.2$.

Figure 1 shows the trajectories for the importance weights $\{\omega_j\}_{j=1}^p$ during the first 50 iterations of the algorithm at the 95% confidence level. In all the considered cases, the trajectories corresponding to terms with non-zero coefficients are clearly distinguished from the others after a few iterations. The inclusion importance for the relevant terms increases to around 1, while the others only have importance weights near 0.5. The graphs show that the MSCS-AS algorithm samples with probability progressively concentrating on the true model terms, while unimportant terms are sampled quite randomly. This behaviour mimics the structure of the true MSCS, $\widehat{\Gamma}_\alpha$, thus enabling us to detect MSCS models at

Table 2. Monte Carlo estimates of MSCS coverage probability and cardinality by exhaustive search for Logistic regression (Model 3) and Poisson regression (Model 4) under varying confidence level $1 - \alpha$, sample size n, and number of variables p. Results based on 500 MC runs.

		Setting 1				Setting 2			
	$n =$	100		250		100		250	
	$p =$	8	12	8	12	8	12	8	12
	α				Model 3				
Coverage(%)	0.10	86.6	80.0	86.6	87.0	86.6	83.6	87.4	89.2
	0.05	92.4	89.6	94.6	91.2	93.0	91.8	93.3	94.6
	0.01	97.6	97.4	99.2	99.0	96.6	97.6	98.6	98.6
Cardinality	0.10	17.5	97.3	14.3	56.8	35.4	459.5	18.2	195.3
	0.05	20.8	147.8	15.2	60.9	43.6	612.6	21.5	252.4
	0.01	32.4	257.0	15.9	67.0	62.0	941.4	29.2	383.6
					Model 4				
Coverage(%)	0.10	89.0	90.6	90.4	89.8	90.6	89.2	89.0	90.4
	0.05	95.6	94.2	95.2	94.6	95.2	94.2	94.6	95.0
	0.01	99.4	99.4	99.0	99.0	99.8	99.4	98.6	99.0
Cardinality	0.10	85.2	871.6	24.6	148.7	75.6	1,269.3	47.9	597.8
	0.05	109.6	1,217.8	32.4	210.3	94.1	1,633.6	51.0	796.7
	0.01	157.9	1,987.0	54.8	402.1	124.9	2,177.4	71.3	1,167.7

a much cheaper computational cost than an exhaustive search of Γ.

As suggested in Section 3 (Step 5 of the algorithm), we stop updating the importance weights shortly after the sequence of significance levels $\alpha^{(0)}, \alpha^{(1)}, \ldots$ reaches the target significance level α^*. However, after $\alpha^{(t)}$ reaches α^*, the weights of the irrelevant predictors eventually converge to 0 or 1 according to Kolmogorov's zero-one law, thus one should stop before that happens. For example, stopping shortly after $\alpha^{(t)} = \alpha^* = 0.05$ – say around 15 iterations – already enables us to detect useful predictors from the rest.

To illustrate that the MSCS-AS algorithm generates MSCS models with large probability, we stopped at iteration 15 and sampled 10^6 models using $p(\cdot; \hat{\boldsymbol{\omega}}^{(15)})$, where $\hat{\boldsymbol{\omega}}^{(15)}$ is the importance weight at the 15th iteration. For the binomial regression model with $p = 100$ predictors, 80.5% of the models generated were included in MSCS. For the Poisson regression model with $p = 100$ predictors, 76.6% of the total models generated were in the MSCS. In comparison, if we generated 10^6 models using the uninformative weights $\boldsymbol{\omega}^{(T)} = (0.5, \ldots, 0.5)$, the proportion of MSCS models is basically 0, due to the largeness of the model space.
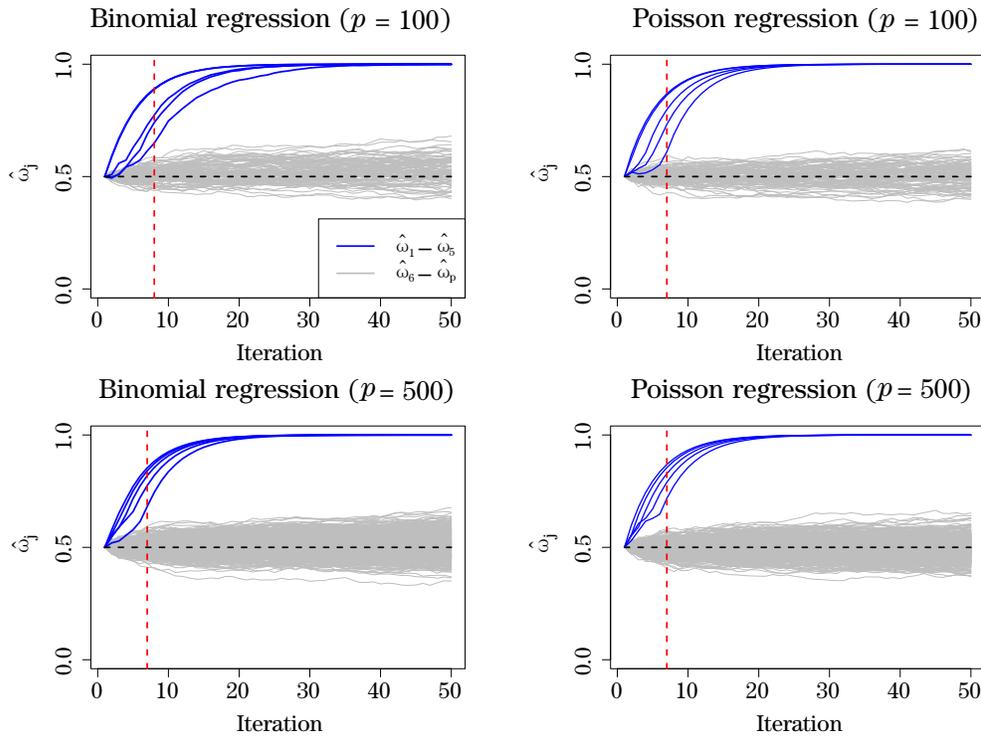
Figure 1. Sampling weights in $p(\cdot;\boldsymbol{\omega})$ for 50 iterations of the MSCS-AS algorithm at the 95% confidence level. The vertical dashed line corresponds to iteration t such that $\alpha^{(t)} = \alpha^*$. Left and right panels correspond to Binomial and Poisson regression models described in Section 4. Settings for the algorithm: $B = 300$, $\zeta = 0.25$, $\alpha^* = 0.05$, and $\xi = 0.2$.

5. Real Data Examples

Example 1: European Escherichia coli(E.coli) O104:H4 outbreak data. We applied the MSCS methodology to the E.coli data as described in Edwards, Pope and Holt (2016). E.coli O104:H4 is a particularly aggressive pathogen and caused a serious outbreak in northern Germany in 2011 (Rasko et al. (2011)). Both during and after the outbreak, scientists examined the genome of E.coli to find genetic causes for the severity of the outbreak. The data set used here consists of 56 outbreak isolates. For each isolate, 10 genes (or hypothetical genes) in the O104:H4 pangenome (a full collection of genes in a species of bacteria) that have been identified as might have been associated with the outbreak are considered. The main goal of our analysis was to select a model that could explain the most meaningful interaction effects between those genes.

The presence of genes in E.coli is denoted by binary variables taking values 1 when the correspondent gene is present and 0 otherwise. Let $\mathbf{Y} = (Y_1, \ldots, Y_{10})$ be a random variable with 10 binary variables each denoting the activity of a particular gene. The pmf of $\mathbf{Y}$ is modelled by the Ising Model,

$$P(\mathbf{Y} = \boldsymbol{y}; \boldsymbol{\theta}) = \exp\left(\sum_{1 \leq j \leq k \leq 10} \theta_{j,k} y_j y_k + \psi(\boldsymbol{\theta})\right), \tag{5.1}$$

where $\boldsymbol{\theta} = (\theta_{j,k})_{1 \leq j \leq k \leq 10}$ is the parameter of interest with $p = 55$ and $\psi(\boldsymbol{\theta})$ is the normalizing constant. The variable $\theta_{j,j}$ is regarded as the main effect for gene j, whilst $\theta_{j,k}$ is interpreted as an interaction effect between genes j and k. Here, we wish to choose $\theta_{j,k} \neq 0$ if genes j and k have interaction (in the same group) and $\theta_{j,k} = 0$ otherwise. In our analysis, we assume $\theta_{j,j} \neq 0$ for $1 \leq j \leq 10$, meaning that the main effects are always included. The total number of possible models in Γ is $B_{10} = 115{,}975$, corresponding to the 10-th Bell number, which counts the number of different ways to partition a set containing 10 elements.

Here we used exhaustive search to construct the MSCS for $\boldsymbol{\theta}$. The MSCSs at the 90, 95 and 99% confidence levels contained 7, 12 and 38 models, respectively; these numbers are small compared to the model space size, meaning that most of the models in Γ are rejected by the LRT procedure and there is not too much model selection uncertainty here. The jk-th element of the matrix in Figure 2 represents the inclusion importance (II) for the variable $\theta_{j,k}$. Certain gene pairs have high II values; for example, pairwise interactions among genes 1141, 3614, 5245, 0051, 0904, 0901 are close to 1 in terms of II values, suggesting that such genes form a synergetic network associated with the outbreak occurrence. Other genes, such as 6065, 0233, 4619, 4633, show inclusion importance for interaction effects close or smaller than 0.5, which suggests that the corresponding interactions are small or irrelevant.

Example 2: Australian breast cancer family study data. We applied the MSCS methodology to the ABCFS genotype data, consisting of 356 observations (284 breast cancer patients and 72 controls). Cases were obtained from the Australian Breast Cancer Family Study (ABCFS) (Dite et al. (2003)), while controls were from the Australian Mammographic Density Twins and Sisters Study by Odefrey et al. (2010). Patients were genotyped using a Human610-Quad beadchip array. The response was the binary disease status (presence/absence of breast cancer), while the predictors were 50 SNPs, measured at different loci encoding a candidate susceptibility pathway (probe IDs are listed in Figure 3. To model the binary disease status, we used a logistic regression model. MSCS
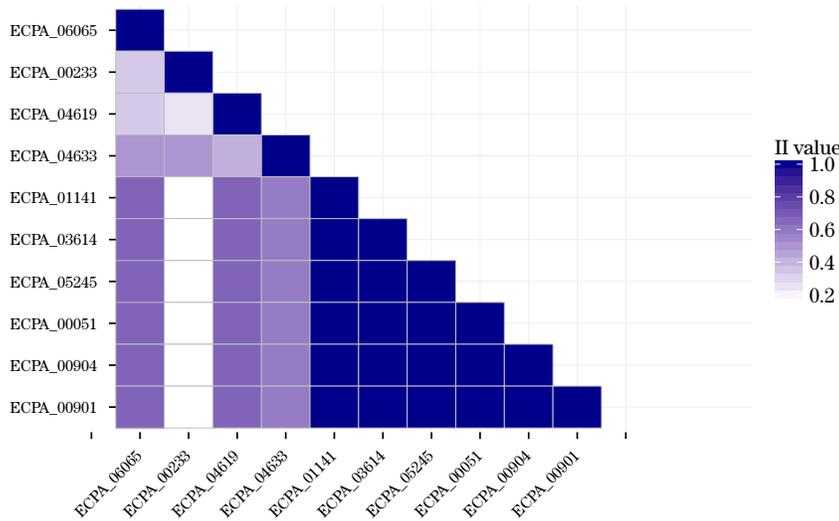
Figure 2. Inclusion Importance (II) for $\theta_{j,k}$ in model (5.1). The II values are calculated from the 95%-MSCS.

models were sampled using the MSCS-AS algorithm described in Section 3 with tuning parameters $\alpha^* = 0.05$, $\zeta = 0.25$, $B = 300$ and $B^{(T)} = 10^6$.

Figure 3 (top) shows the models selected by forward step-wise AIC and BIC (F-AIC and F-BIC), and penalized likelihood methods under Lasso (Tibshirani (1996)), SCAD (Fan and Li (2001)) and MCP (Zhang (2010)) penalties where tuning parameters were chosen by five-fold cross-validation. Figure 3 (bottom) shows Inclusion Importance, II values, for each SNP. The vertical bars represent the 95% bootstrap confidence intervals for II values. Here SNPs with high II values show considerable overlap with those selected by the other methods. Particularly, most of the SNPs reported as important by more than one model selection method have large II values. Predictors with II values that are not significantly larger than 0.5 may not necessarily be important, and need to be consider more carefully.

The p-values of the F-test for the AIC, BIC, Lasso, SCAD and MCP models were 0.96, 0.09, 0.88, 0.86 and 0.80. respectively. Therefore, while the BIC model is included in the MSCS at the 95% and 99% levels, it is not accepted at the 90% confidence level. This sugggests that F-BIC model misses some important SNP predictors and is not as reliable as other models. Moreover, there are only 5 SNPs (rs0082248_A, rs12470143_T, rs2754530_T, rs8192207_T, rs2257157_G) that have relatively large II values and confidence intervals significantly above 0.5. Many
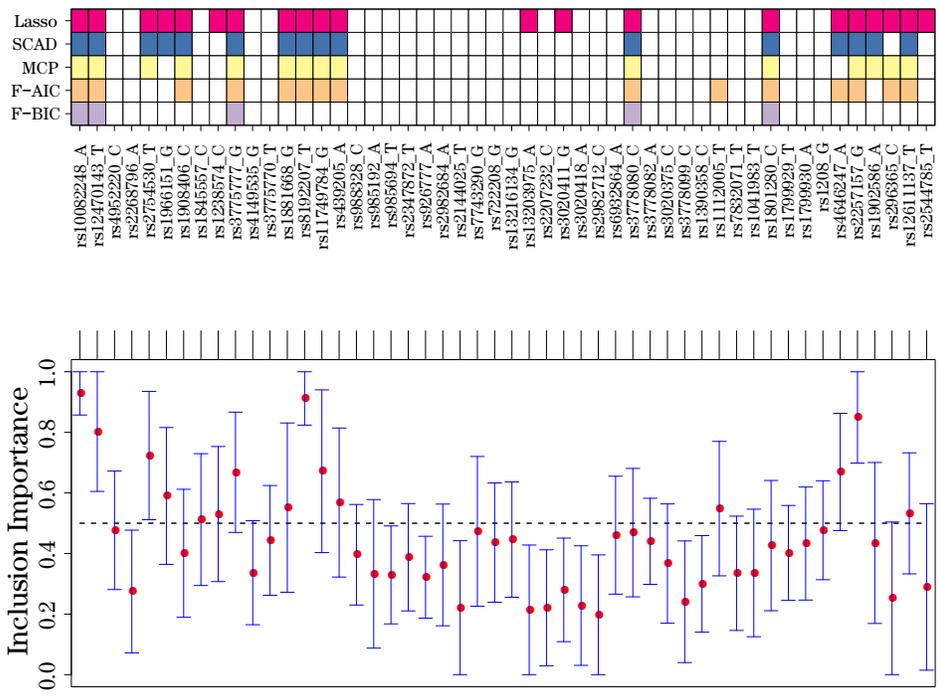
Figure 3. Analysis of the ABCFS case-control genotype data. Top: model selection by penalized likelihood methods with Lasso, SCAD and MCP penalties and step-wise forward AIC and forward BIC. Darken cells denote selected variables. Bottom: Estimate of II by MSCS-AS algorithm(mark points) and the 95% bootstrap confidence intervals. The number of bootstrap replicates is 50. Settings for the MSCS-AS algorithm parameters: $B = 300$, $B^{(T)} = 10^6$, $\alpha^* = 0.05, \zeta = 0.25, \xi = 0.2$.

.

predictors chosen by some of the methods have II confidence intervals covering 0.5. This suggests that the sample is not sufficiently informative to declare such terms relevant so they should be further studied with particular care. To confirm this, we investigated the marginal significance of all the selected SNPs in each of those models. At the 0.05 level, we have 8 (Lasso, SCAD and MCP) to 12 (AIC) significant coefficients, while the 5 SNPs with II interval not including 0.5 are significant in all cases.

Finally, we looked at the instability of common model selection methods for this dataset. We considered the selected models as in Figure 3 (top). For each model we obtained the fitted values $\{\hat{Y}_i\}_{i=1}^n$, equal to $\{\hat{p}\}_{i=1}^n$, the estimated probability in the logistic regression. Next, parametric bootstrap was used to generate bootstrap replicates $\{X_i, Y_i^*\}_{i=1}^n$, and all the methods were applied again

Table 3. The number of SNPs in the original selected model and the average Hamming distance between the 500 bootstrapped models and the original selected model.

	Lasso	SCAD	MCP	F-AIC	F-BIC
Size	21	16	15	15	5
AHD	14.74	11.132	10.44	11.58	5.59

to corresponding bootstrap samples. We repeated this step $S = 500$ times and computed the average Hamming distances (AHD) of the S bootstrapped Lasso, SCAD, MCP, F-AIC and F-BIC models to the the respective models obtained from the original sample, see Table 3. The AHD is large compared the size of the models, which means for each bootstrapped sample, those model selection methods choose quite different predictors.

6. Conclusion and Final Remarks

Our methodology introduces new tools supporting the activity of model selection in the context of likelihood-based inference. Since the MSCS is asymptotically guaranteed to contain the true model at a pre-specified confidence level, it represents a natural extension of the familiar notion of confidence intervals to the model selection framework. Our Theorem 4 suggests that important variables tend to appear in the MSCS models with large probability as $n \to \infty$, while unimportant terms appear randomly with frequency not significantly larger than 0.5. By looking at the variables appearing frequently in the MSCS one can also choose a single central model representing the entire MSCS by taking predictors with inclusion importance significantly larger than 0.5. In the future, developing a theoretical understanding of the optimal way to combine MSCS models would be valuable as it can potentially lead to improved model combining and model selection strategies.

The main focus of the current MSCS approach is based on maximum likelihood estimation. We have shown that in exponential family models this requires $p = o(n^{2/3})$ and correct model specification for the MSCS to be meaningful. In the future, higher-dimensional problems can be pursued, by replacing the LRT statistics with other tools to construct the MSCS, e.g., using penalized likelihood methods. Computational methods to tackle the case where the model space and MSCS is large is also of interest.

Acknowledgements

Appendix: Proofs

The quantities such as p, $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\Theta}$ may depend on n, hence array asymptotics are considered in this section.

Proof of Theorem 2.5. Since within the considered exponential family affine mappings are preserved, without loss of generality we assume $E(\mathbf{Y}_i) = A'(\boldsymbol{\theta}^*) = 0$ and $\mathrm{cov}(\mathbf{Y}_i) = A''(\boldsymbol{\theta}^*) = I_p$.

Let $\mathbf{V}_{\boldsymbol{\theta}} = \mathbf{Y}_{\boldsymbol{\theta}} - E(\mathbf{Y}_{\boldsymbol{\theta}})$, where $\mathbf{Y}_{\boldsymbol{\theta}} \sim f(\boldsymbol{y}; \boldsymbol{\theta})$. We need some regularity conditions for the exponential family model $f(; \boldsymbol{\theta})$.

(A1) $EY_{ij}^6 < \infty$, $\quad (j = 1, \ldots, p)$.
(A2) $\sup_{\substack{\|a\|=1 \\ \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2 \lesssim p/n}} \left| E(a^T \mathbf{V}_{\boldsymbol{\theta}})^3 \right| \lesssim \sqrt{n/p}$, and $\sup_{\|a\|=1} \left| E(a^T \mathbf{V}_{\boldsymbol{\theta}})^3 \right| = O(1)$.
(A3) $\sup_{\substack{\|a\|=1 \\ \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2 \lesssim p/n}} E(a^T \mathbf{V}_{\boldsymbol{\theta}})^4 = O(1)$.

Suppose we have $\boldsymbol{\gamma}$ which is a model not containing all the elements in $\boldsymbol{\gamma}^*$. Let $\overline{\mathbf{Y}} = \sum_{i=1}^{n} \mathbf{Y}_i/n$ and $\overline{\mathbf{Y}}_{\boldsymbol{\gamma}} = \sum_{i=1}^{n} \mathbf{Y}_{i,\boldsymbol{\gamma}}/n$, where $\mathbf{Y}_{i,\boldsymbol{\gamma}}$ denote the vector with elements equal to $\mathbf{Y}_i$ for indexes in $\boldsymbol{\gamma}$ and zero otherwise.

Let $\mathbf{G} = \boldsymbol{\theta}^* - \boldsymbol{\theta}_{\boldsymbol{\gamma}}^*$ satisfy $\|\mathbf{G}\| \gtrsim \sqrt{p/n}$ and assume $p = o(n^{2/3})$, where $\boldsymbol{\theta}_{\boldsymbol{\gamma}}^*$ is the $p_{\boldsymbol{\gamma}^*}$-variate vector with components equal to $\boldsymbol{\theta}^*$ at $\boldsymbol{\gamma} \cap \boldsymbol{\gamma}^*$ and equal to zero otherwise.

Lemma A1. *If* (A.2) *and* (A.3) *hold, for the MLEs*, $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$, *we have* $\|\boldsymbol{\theta}_{\boldsymbol{\gamma}} - \boldsymbol{\theta}_{\boldsymbol{\gamma}}^*\| = O_p\left(\sqrt{p_{\boldsymbol{\gamma}}/n}\right)$, *and* $\|\boldsymbol{\theta}_{\boldsymbol{\gamma}} - \boldsymbol{\theta}_{\boldsymbol{\gamma}}^* - \overline{\mathbf{Y}}_{\boldsymbol{\gamma}}\| = O_p\left(p_{\boldsymbol{\gamma}}/n\right)$.

Lemma 1 combines Theorems 2.1 and 3.1 in Portnoy (1988). The existence of MLE for $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ in a L-2 neighbourhood of order $\sqrt{p_{\boldsymbol{\gamma}}/n}$ is still valid by simply changing the true model $\boldsymbol{\theta}^*$ to the partial model $\boldsymbol{\theta}_{\boldsymbol{\gamma}}^*$.

Lemma A2. *If* (A.2) *and* (A.3) *hold, for model* $\boldsymbol{\gamma}$, $\left|A(\boldsymbol{\theta}_{\boldsymbol{\gamma}}^*) - A(\boldsymbol{\theta}^*)\right| = \|\boldsymbol{\theta}_{\boldsymbol{\gamma}}^* - \boldsymbol{\theta}^*\|^2/2 + o(\|\boldsymbol{\theta}_{\boldsymbol{\gamma}}^* - \boldsymbol{\theta}^*\|^2)$.

Proof. From (A.3), by Taylor expansion, we have

$$\left|A(\boldsymbol{\theta}_{\boldsymbol{\gamma}}^*) - A(\boldsymbol{\theta}_n^*)\right| = \left| \frac{1}{2} \|\boldsymbol{\theta}_{\boldsymbol{\gamma}}^* - \boldsymbol{\theta}^*\|^2 + \frac{1}{6} E\left\{(\boldsymbol{\theta}_{\boldsymbol{\gamma}}^* - \boldsymbol{\theta}_n^*)^T \mathbf{V}_{\boldsymbol{\theta}^*}\right\}^3 \right| + O(\|\boldsymbol{\theta}_{\boldsymbol{\gamma}}^* - \boldsymbol{\theta}_n^*\|^4).$$

Applying (A.2) gives $E\left\{(\boldsymbol{\theta}_{\boldsymbol{\gamma}}^* - \boldsymbol{\theta}^*)^T \mathbf{V}_{\boldsymbol{\theta}^*}\right\}^3 \lesssim \|\boldsymbol{\theta}_{\boldsymbol{\gamma}}^* - \boldsymbol{\theta}^*\|^3$, which completes the proof.

Lemma A3. *If* (A.1), (A.2) *and* (A.3) *hold, when* $p - p_{\boldsymbol{\gamma}} \to \infty$ *we have*

$$\frac{n\left(\|\overline{\boldsymbol{Y}}\|^2 - \|\overline{\boldsymbol{Y}_{\boldsymbol{\gamma}}}\|^2\right) + 2n\,\boldsymbol{G}^T\overline{\boldsymbol{Y}} - (p - p_{\boldsymbol{\gamma}})}{\sqrt{2(p - p_{\boldsymbol{\gamma}} + 2\delta_n)}} \xrightarrow{\mathscr{D}} \mathcal{N}_1(0, 1), \qquad (A.1)$$

where $\delta_n = n\|\boldsymbol{G}\|^2$.

Proof. Let $\mathbf{T}_k = \sum_{i=1}^k \mathbf{Y}_i$, $\mathbf{T}_{k,\boldsymbol{\gamma}} = \sum_{i=1}^k \mathbf{Y}_{i,\boldsymbol{\gamma}}$ and $S_k = \|\mathbf{T}_k\|^2 - \|\mathbf{T}_{k,\boldsymbol{\gamma}}\|^2 + 2k\mathbf{G}^T\mathbf{T}_k - k(p - p_{\boldsymbol{\gamma}})$. We have $E(S_n) = 0$ and $\mathrm{var}(S_n) = 2n^2(p - p_{\boldsymbol{\gamma}} + 2n\|\mathbf{G}\|^2)$.

Let $D_k = S_k - S_{k-1}$, then $D_k = 2\mathbf{Y}_k^T\mathbf{T}_{k-1} + \|\mathbf{Y}_k\|^2 - 2\mathbf{Y}_{k,\boldsymbol{\gamma}}^T\mathbf{T}_{\boldsymbol{\gamma},k-1} - \|\mathbf{Y}_{k,\boldsymbol{\gamma}}\|^2 + 2\mathbf{G}^T\mathbf{T}_{k-1} + 2k\mathbf{G}^T\mathbf{Y}_k - (p - p_{\boldsymbol{\gamma}})$.

Take $\sigma_k^2 = ED_k^2$ and $s_k^2 = \sum_{i=1}^k \sigma_i^2$. As $\mathbf{Y}_k$ and $\mathbf{T}_{k-1}$ are independent, simple calculation shows that

$$s_n^2 = \sum_{k=1}^n \sigma_k^2 \lesssim n^2 p + n^3\|G\|^2, \qquad (A.2)$$

Let $\mathcal{F}_k = \mathcal{F}(\mathbf{Y}_1, \dots, \mathbf{Y}_k)$ denote the σ-field generated by $\mathbf{Y}_1, \dots, \mathbf{Y}_k$. Then $\{S_k\}$ are martingales on $\{\mathcal{F}_k\}$, and $\{D_k\}$ are the martingale differences. From Chow and Teicher (1978), by the Martingale Central Limit Theorem , we have $S_n/\left(n\sqrt{p - p_{\boldsymbol{\gamma}} + \delta_{\boldsymbol{\gamma}}}\right) \xrightarrow{\mathscr{D}} \mathcal{N}_1(0, 1)$ if

$$\sum_{k=1}^n \frac{E|D_k|^3}{s_n^3} \to 0, \quad \text{and} \quad \sum_{k=1}^n \frac{E|E(D_k^2|\mathcal{F}_{k-1}) - \sigma_k^2|}{s_n^2} \to 0. \qquad (A.3)$$

Since $E(Y_{ij}^6) < \infty$, by Proposition A.3 in Portnoy (1988), we have $E(\mathbf{Y}_k^T\mathbf{T}_{k-1})^6 \lesssim k^3p^3$ and therefore $\sum_{k=1}^n E|D_k|^3 \lesssim n^{5/2}p^{3/2} + n^2p^{5/2} + n^4\|G\|^3$,

Together with Equation (A.2) this implies $\sum_{k=1}^n E|D_k|^3/s_n^3 \to 0$.

Next,

$$\sum_{k=1}^n E|E(D_k^2|\mathcal{F}_{k-1}) - \sigma_k^2| \leq \left[E\left\{E(D_k^2|\mathcal{F}_{k-1}) - \sigma_k^2\}^2\right]^{1/2} \qquad (A.4)$$

$$\lesssim n^{3/2}p^{3/2} + n^2p^{1/2} + n^{3/2}\|G\|^2, \qquad (A.5)$$

So we have $\sum_{k=1}^n E|E(D_k^2|\mathcal{F}_{i-1}) - \sigma_k^2|/s_n^2 \to 0$.

Proof of Theorem 2.5. Under the alternatives, we have

$$\Lambda_{\boldsymbol{\gamma}} = 2n(\boldsymbol{\theta}_{\boldsymbol{\gamma}_f} - \boldsymbol{\theta}_{\boldsymbol{\gamma}})^T\overline{\boldsymbol{Y}} - 2n\left\{A(\boldsymbol{\theta}_{\boldsymbol{\gamma}_f}) - A(\boldsymbol{\theta}_{\boldsymbol{\gamma}})\right\}$$
$$= 2n\left\{(\boldsymbol{\theta}_{\boldsymbol{\gamma}_f} - \boldsymbol{\theta}^*) - (\boldsymbol{\theta}_{\boldsymbol{\gamma}} - \boldsymbol{\theta}_{\boldsymbol{\gamma}}^*)^T\right\}\overline{\boldsymbol{Y}} - 2n\left\{A(\boldsymbol{\theta}_{\boldsymbol{\gamma}_f}) - A(\boldsymbol{\theta}_{\boldsymbol{\gamma}}^*) - A(\boldsymbol{\theta}_{\boldsymbol{\gamma}}) + A(\boldsymbol{\theta}_{\boldsymbol{\gamma}}^*)\right\}$$
$$- 2n\left\{(\boldsymbol{\theta}_{\boldsymbol{\gamma}}^* - \boldsymbol{\theta}^*)^T\overline{\boldsymbol{Y}} - A(\boldsymbol{\theta}_{\boldsymbol{\gamma}}^*) + A(\boldsymbol{\theta}^*)\right\}. \qquad (A.6)$$

The first two terms of the LRT statistic in (A.6) are approximately $n(\|\overline{\mathbf{Y}}\|^2 - \|\overline{\mathbf{Y}}_{\boldsymbol{\gamma}}\|^2) + O_p(p^2/n)$. Lemma 2 implies that the third term in (A.6) is approx-

imately equal to $-2n(\boldsymbol{\theta}_{\boldsymbol{\gamma}}^* - \boldsymbol{\theta}^*)\overline{\mathbf{Y}} + n\|\boldsymbol{\theta}_{\boldsymbol{\gamma}}^* - \boldsymbol{\theta}^*\|^2$. Hence, applying Lemma 3 completes the proof.

Proof of Theorem 2. We show that if the sufficient condition in Theorem 2 is satisfied, we have

$$P\left(\bigcup_{\boldsymbol{\gamma}\in\Gamma_u} \{\Lambda_{\boldsymbol{\gamma}} \le q(\alpha; d_{\boldsymbol{\gamma}})\}\right) \to 0, \quad \text{as } n \to \infty.$$

Then

$$P\left(\bigcup_{\boldsymbol{\gamma}\in\Gamma_u} \{\Lambda_{\boldsymbol{\gamma}} \le q(\alpha; d_{\boldsymbol{\gamma}})\}\right) \le \sum_{k=1}^{p-1} \sum_{\substack{d_{\boldsymbol{\gamma}}=k \\ \boldsymbol{\gamma}\in\Gamma_u}} P\big(\Lambda_{\boldsymbol{\gamma}} \le q(\alpha; k)\big) \tag{A.7}$$

$$\le \sum_{k=1}^{p-1} \exp\left(K_n(k)\right) \max_{\substack{d_{\boldsymbol{\gamma}}=k \\ \boldsymbol{\gamma}\in\Gamma_u}} P\left(\Lambda_{\boldsymbol{\gamma}} \le q(\alpha; k)\right). \tag{A.8}$$

We assume the exponential bound as in (2.6). Thus, combining Lemma 8.1 in Birgé (2001) and Theorem A in Inglot (2010),

$$\max_{\substack{d_{\boldsymbol{\gamma}}=k \\ \boldsymbol{\gamma}\in\Gamma_u}} P\left(\Lambda_{\boldsymbol{\gamma}}(d_{\boldsymbol{\gamma}}, \delta_{\boldsymbol{\gamma}}) \le q(\alpha; k)\right) \le \max_{\substack{d_{\boldsymbol{\gamma}}=k \\ \boldsymbol{\gamma}\in\Gamma_u}} P\left(X_{\delta_{\boldsymbol{\gamma}}, k} \le q(\alpha, k)\right) + c_1 \exp\left[-c_2 K_n(k)\right]$$

$$\le \exp\left(-\min_{\substack{d_{\boldsymbol{\gamma}}=k \\ \boldsymbol{\gamma}\in\Gamma_u}} \frac{\left(\delta_{\boldsymbol{\gamma}} + 2\log(\alpha) - 2\sqrt{-k\log(\alpha)}\right)^2}{2(k + 2\delta_{\boldsymbol{\gamma}})}\right)$$

$$+ c_1 \exp\left(-c_2 K_n(k)\right) \tag{A.9}$$

where c_1, c_2 are positive constants. When $\min_{d_{\boldsymbol{\gamma}}=k, \boldsymbol{\gamma}\in\Gamma_u} \delta_{\boldsymbol{\gamma}}/K_n(k) > B$ for some large enough positive constant B, the first term in the upper bound (A.9) will be smaller than $\exp(-(c_3 - 1)K_n(k))$ for some $c_3 > 1$. Therefore, we have

$$P\left(\bigcup_{\boldsymbol{\gamma}\in\Gamma_u} \{\Lambda_{\boldsymbol{\gamma}} \le q(\alpha; d_{\boldsymbol{\gamma}})\}\right) \lesssim \sum_{k=1}^{p-1} \exp(-\min(c_2, c_3 - 1)K_n(k)) \to 0. \tag{A.10}$$

Proof of Theorem 3. Let $\hat{\boldsymbol{\gamma}}$ be the largest model that are nested in all the models in MSCS. It may be the intercept only model. With probability at least $1-\alpha$ the true model $\boldsymbol{\gamma}^*$ is included in the MSCS, $P\big(\boldsymbol{\gamma}^* \in \widehat{\Gamma}_\alpha\big) \ge 1-\alpha$. Therefore

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}} P_{\boldsymbol{\beta}}\left(\hat{\boldsymbol{\gamma}} \ne \boldsymbol{\gamma}^*\right) \le \sup_{\boldsymbol{\beta}\in\mathcal{B}} P_{\boldsymbol{\beta}}(\boldsymbol{\gamma}^* \notin \widehat{\Gamma}_\alpha) + \sup_{\boldsymbol{\beta}\in\mathcal{B}} P_{\boldsymbol{\beta}}(\mathcal{D}^C) \tag{A.11}$$

$$\le \alpha + 1 - \inf_{\boldsymbol{\beta}\in\mathcal{B}} P_{\boldsymbol{\beta}}(\mathcal{D}). \tag{A.12}$$

It follows that if $\inf_{\boldsymbol{\beta}\in\mathcal{B}} P_{\boldsymbol{\beta}}(\mathcal{D}) \to 1$ for any subsequence of $\{n_j\} \subset \{1, 2, \ldots\}$,

we must have $\sup_{\boldsymbol{\beta} \in \mathcal{B}} P_{\boldsymbol{\beta}} (\hat{\boldsymbol{\gamma}} \neq \boldsymbol{\gamma}^*) \leq \alpha'$ for some $\alpha < \alpha' < 1/2$ when n_j is large enough in the subsequence. Hence if we can show $\sup_{\boldsymbol{\beta} \in \mathcal{B}} P_{\boldsymbol{\beta}} (\hat{\boldsymbol{\gamma}} \neq \boldsymbol{\gamma}^*) \geq 1/2$ for a small enough $c > 0$ in the definition of $\mathcal{B}$, the theorem is proved.

This is a traditional minimax framework where Fano's inequality can be applied. Without loss of generality, we assume $\sigma^2 = 1$ for the error variance. Consider a packing set N_{ϵ_n} in $\mathcal{B} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_0 = r^*$ and $\|f_{\boldsymbol{\beta}}\|_n^2 \leq cK_n(r^*)\}$ with packing distance ϵ_n, a small fraction of $\sqrt{cK_n(r^*)}$ under the $\|\|_n$ norm. Let $\boldsymbol{\beta}$ be uniformly choosen from N_{ϵ_n}. Using similar arguments as in the proof of Theorem 11 in Wang et al. (2014), under the SRC, by choosing the constant c small enough, the mutual information $\mathbf{I}(\boldsymbol{\beta}; \{X_i, Y_i\}_{i=1}^n)$ between the random $\boldsymbol{\beta}$ and the observations is upper bounded by $\delta = cK_n(r^*)/2$, and the local packing ϵ_n-entropy $\log |N_{\epsilon_n}|$ is lower bounded by $2\delta + 2\log 2$. Apply Fano's inequality (see, e.g., Yang and Barron (1999)) to this linear regression model gives

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}} P_{\boldsymbol{\beta}} (\hat{\boldsymbol{\gamma}} \neq \boldsymbol{\gamma}^*) \geq 1 - \frac{\mathbf{I}(\boldsymbol{\beta}; \{X_i, Y_i\}_{i=1}^n) + \log 2}{\log |N_{\epsilon_n}|} \geq \frac{1}{2}, \qquad (A.13)$$

and completes the proof.

Proof of Theorem 4. (i) Let $\boldsymbol{\gamma}$ be a model missing at least one variable in $\boldsymbol{\gamma}^*$. From Theorem 2, $P(\boldsymbol{\gamma} \in \widehat{\Gamma}_\alpha) \to 0$, as $n \to \infty$. Hence, for all θ_j in $\boldsymbol{\theta}^*$, $\lim_{n \to \infty} P(II_j = 1) = 1$. This completes the first part of the theorem.

(ii) Let $\widetilde{\Gamma} = \{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_N\}$ be the set of models larger than $\boldsymbol{\gamma}^*$. The construction of MSCS implies $\lim_{n \to \infty} P(\boldsymbol{\gamma}_i \in \widehat{\Gamma}_\alpha) \geq 1 - \alpha$ for any $1 \leq i \leq N$. Let X_i be the random variable taking value 1 when $\boldsymbol{\gamma}_i \in \widehat{\Gamma}_\alpha$ and 0 otherwise. When $k \notin \boldsymbol{\gamma}^*$, there are $N/2$ models in $\widetilde{\Gamma}$ that contain θ_k. Without loss of generality, let $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_{N/2}$ be models containing k. Let $Y_1 = \sum_{i=1}^{N/2} X_i$, denoting the number of models in $\{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_{N/2}\}$ that are included in $\widehat{\Gamma}_\alpha$. Similarly, let $Y_2 = \sum_{i=N/2+1}^N X_i$. Then $Y_1 \leq N/2$ and $E(Y_2) \geq N(1 - \alpha)/2$. The II_k defined in Section 2.4 is $Y_1/(Y_1 + Y_2)$ with probability going to 1 as $n \to \infty$. For any $0 < \Delta \leq 1/2$ we have

$$\begin{aligned}
P\left(\frac{Y_1}{Y_1 + Y_2} \geq \frac{1}{2} + \Delta\right) &\leq P\left(\frac{N/2}{N/2 + Y_2} \geq \frac{1}{2} + \Delta\right) \\
&= P\left(N/2 - Y_2 \geq \frac{2\Delta N}{1 + 2\Delta}\right) \\
&\leq \frac{\{N/2 - E(Y_2)\}(1 + 2\Delta)}{2N\Delta} \\
&\leq \frac{\alpha(1 + 2\Delta)}{4\Delta},
\end{aligned}$$

where the second inequality follows from Markov's inequality. This completes the proof.

References

Birgé, L. (2001). *State of the Art in Probability and Statistics.* IMS Lecture Notes **37**, 113–133.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, New York.

Chow, Y. S. and Teicher, H. (1978). *Probability Theory: Independence, Interchangeability, Martingales.* Springer, New York.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging.* Cambridge University Press, Cambridge.

Costa, A., Jones, O. D. and Kroese, D. (2007). Convergence properties of the cross-entropy method for discrete optimization. *Oper. Res. Lett.* **35**, 573–580.

Dite, G. S., Jenkins, M.K., Southey, M. C., Hocking, J. S., Giles, G. G., McCredie, M. R. E., Venter, D. J. and Hopper, J. L. (2003). Familial risks, early-onset breast cancer and brca1 and brca2 germline mutations. *J. Natl. Cancer Inst.* **95**, 448–457.

Edwards, D., Pope, B. J. and Holt, K. E. (2016). Reddog tutorial. GitHub techinical report, Web: `https://github.com/katholt/RedDog/blob/master/docs/RedDogTutorial.pdf`.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Trans. Inform. Theory.* **57**, 5467–5484.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928–961.

Fan, J. and Zhou, W. X. (2015). Guarding from spurious discoveries in high dimension. *J. Mach. Learn. Res.* **17**, 1–34.

Ferrari, D. and Yang, Y. (2015). Confidence sets for model selection by F-testing. *Statist. Sinica.* **25**, 1637-1658.

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

Inglot, T. (2010). Inequalities for quantiles of the chi-square distribution. *Probab. Math. Statist.* **30**, 339–351.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses.* 3rd edition. Springer, New York.

Mcquarrie, A. D. and Tsai, C. L. (1998). *Regression and Time Series Model Selection.* World Scientific, Singapore.

Murphy, S. A. (1993). Testing for a time dependent coefficient in cox regression model. *Scand. J. Stat.* **20**, 35–50.

Odefrey, F., Stone, J., Gurrin, L.C., Byrnes, G. B., Apicella, C.,Dite, G. S., Cawson, J. N., Giles, G. G., Treloar, S. A., English, D. R., Hopper, J. L. and Southey, M. C. (2010). Common genetic variants associated with breast cancer and mammographic density measures that predict disease. *Cancer Res.* **70**, 1449–1458.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when

the number of parameters tends to infinity. *Ann. Statist.* **16**, 356–366.

Rasko, D. et al. (2011). Origins of the e. coli strain causing an outbreak of hemolytic-uremic syndrome in germany. *N. Engl. J. Med.* **365**, 709–717.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.* **100**, 94–108.

Rubinstein, R. Y. and Kroses, D. P. (2004). *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning.* Springer-Verlag, New York.

Shimodaira, H. (1998). An application of multiple comparison techniques to model selection. *Ann. Inst. Statist. Math.* **50**, 1–13.

Spokoiny, V. (2012). Parametric estimation. finite sample theory. *Ann. Statist.* **40**, 2877–2909.

Spokoiny, V. (2013). Bernstein-von mises theorem for growing parameter dimension. Available at *arXiv:1302.3430.*

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 267–288.

Van der Vaart, A. W. (2000). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

Wang, Z., Paterlini, S., Gao, F. and Yang, Y. (2014). Adaptive minimax regression estimation over sparse l_q-hulls. *J. Mach. Learn. Res.* **15**, 1675–1711.

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27**, 1564–1599.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.

Department of Mathematics and Statistics, Lancaster University.

E-mail: c.zheng5@lancaster.ac.uk

Department of Economics and Management, Free University of Bozen-Bolzano, 39100, Bolzano, Italy.

E-mail: davferrari@unibz.it

School of Statistics, University of Minnesota.

E-mail: yyang@stat.umn.edu