Statistica Sinica **25** (2015), 1637-1658
doi:http://dx.doi.org/10.5705/ss.2014.110

CONFIDENCE SETS FOR MODEL SELECTION BY F-TESTING

Davide Ferrari and Yuhong Yang

University of Melbourne and University of Minnesota

Abstract: We introduce the notion of variable selection confidence set (VSCS) for linear regression based on F-testing. Our method identifies the most important variables in a principled way that goes beyond simply trusting the single winner based on a model selection criterion. The VSCS extends the usual notion of confidence intervals to the variable selection problem: A VSCS is a set of regression models that contains the true model with a given level of confidence. Although the size of the VSCS properly reflects the model selection uncertainty, without specific assumptions on the true model, the VSCS is typically rather large (unless the number of predictors is small). As a solution, we advocate special attention to the set of lower boundary models (LBMs), which are the most parsimonious models not statistically significantly inferior to the full model at a given confidence level. Based on the LBMs, variable importance and measures of co-appearance importance of predictors can be naturally defined.

Key words and phrases: Confidence set, linear regression, model selection, variable selection.

1. Introduction

A statistical model can be interpreted as a story about how the data might have been generated by a particular random process. In many empirical analyses, a relevant question is: “Which story is the most plausible?”. Sometimes, we are in the fortunate situation where the data strongly support one story, and so the corresponding model may be properly singled out as the “truth” for most purposes. More often than not, however, while we wish to select a single model, the data do not clearly support a unique model.

In the literature of model selection, this issue is sometimes referred to as model selection uncertainty (Chatfield (1995), Draper (1995), Hoeting et al. (1999), Yuan and Yang (2005)). A wealth of methods is available in the literature of statistics and machine learning for variable selection. However, often it is difficult to declare a single model as superior to all possible competing models or even among the best set of models, due to the prevailing effect of model selection uncertainty. The methodology proposed in this paper is not meant to compete with existing model selection methods. Rather, it aims to characterize the intrinsic model selection uncertainty associated with the data at hand and to provide information on variable importance that goes beyond the standard single-final-model approach.

A well-established way to address model selection uncertainty is model averaging. It is now well understood that by weighting the candidate models properly, estimation or prediction can be much improved. See, e.g., Hoeting et al. (1999); Yang (2001); Hjort and Claeskens (2003) and references therein. In our view, a fundamental drawback from selecting a single model, which is not sufficiently dealt with by model averaging, is that when a single set of variables is chosen, a wealth of information is possibly thrown away in three key aspects. One is that alternative stories, possibly equally well supported, are ignored, which may be highly undesirable in terms of scientific understanding of the nature of the data. The second aspect is that it does not give any indication of how reliable the selected model is, since uncertainty measures such as standard errors and confidence intervals based on the final model can be highly misleading. The third issue is that centering on a single model alone fails to provide trustworthy association among the predictors in jointly influencing the response variable.

This paper approaches variable selection from a different perspective by reducing the set of all possible collections of the variables to a smaller set, variable selection confidence set (VSCS), that contains the true model with a given level of confidence. Our methodology reflects variable selection uncertainty: if the data are uninformative, distinguishing between models is difficult and the VSCS may contain a large number of interesting models; in the presence of abundant information, the VSCS tends to be much smaller and essentially gives out the true model when the sample size grows to infinity. We begin with a set of predictors of size smaller than the sample size, possibly after a variable screening or an initial variable selection. We then construct an exact VSCS based on F-tests. Such a confidence set can be very large but, as will be seen, some sparse model selection methods sometimes produce a model not in the VSCS, in which case one can be confident that the selected model is too sparse. Next, an important subset of the VSCS is identified, the lower boundary models (LBMs), defined as the smallest models that are not statistically significantly inferior to the full model at a given confidence level. Dropping any term(s) in the LBMs would make the reduced model unfit from a hypothesis testing perspective. At the given confidence level, each model in the LBM set tells a well-justified, most parsimonious story. We show that the LBMs contain the information on the true predictors as $n \to \infty$; at the same time the LBMs are computationally more tractable. The set of LBMs can provide useful information on how many plausible stories are there to explain the data, which predictors are definitely needed in most of the stories and which predictors co-star in most of the stories.

The idea of a confidence set for models has been explored. For example, Shimodaira (1998) advocates the use of a set of models that have AIC values close to the smallest among the candidates based on hypothesis testing. An important work of Hansen, Lunde and Nason (2011) proposes a notion of model confidence set in a framework that does not directly require the specification of the data generating model. The approach is analogous to some step-down procedures for multiple hypothesis testing (e.g., Dudoit, Shaffer and Boldrick (2003), Lehmann and Romano (2006) or Romano and Wolf (2005)), as is mentioned in Hansen, Lunde and Nason (2011). Although we share the same general motivation, our approach is different: we start with a strong linear model assumption and concretely build variable selection confidence sets. The focused framework offers a number of advantages: we achieve exactly the specified coverage probability for the globally optimal model; in our setting the number of predictors, p, is allowed to grow with the sample size, n; in addition to the confidence sets, our approach leads to tools to assess model selection uncertainty.

The rest of the paper is organized as follows. In Section 2, we present the exact VSCS based on F-tests. In Section 3, the subset of LBMs is defined, and the properties of LBMs and variable importance measures are given. In Section 4, we illustrate the utility of our methods based on two data sets. Simulation results are in Section 5. A discussion of the closely related work of Hansen, Lunde and Nason (2011) is in Section 6, followed by final remarks in Section 7. The proofs of the main theorems are deferred to a separate Appendix available on-line.

2. Exact Confidence Set

2.1. Setup

In this section, we construct an exact confidence set (ECS) in terms of coverage probability and study a related issue of detectability of the terms in the true model. Throughout, we assume a normal regression model for the response variable:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{j,i} + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where ϵ_i are i.i.d. from $N(0, \sigma^2)$, for some $\sigma^2 > 0$. The predictors are considered to be fixed and the intercept is always included. Some of the coefficients in $(\beta_1, \ldots, \beta_p)$ are possibly zero. Let γ^* denote the set of indexes of all non-zero terms in the true mean expression. Our main interest is to construct a confidence set of models, $\widehat{\Gamma}$, such that $P(\gamma^* \in \widehat{\Gamma}) \geq 1 - \alpha$, for a given $0 < \alpha < 1$. We use γ as the index of a model, which corresponds to a collection of predictors.

2.2. Exact confidence sets based on F-testing

We use the familiar F-test to look for models that can plausibly be the true model. The full model is assumed to be uniquely fitted by least squares, which is typically appropriate when p is smaller than n. When p is larger than n, screening methods may be needed to reduce the number of predictors to be less than n. When one uses the VSCS to assist a model selection method by providing additional information, a "full model" can be constructed to be a super model of the presently selected model by the method (see Section 4.2). Let γ_f denote the full model. The F-test compares the candidate model γ to the full model γ_f. Particularly, γ is rejected when

$$\widehat{F}(\gamma_f,\gamma)=\frac{\left(RSS_{\gamma}-RSS_{\gamma_f}\right)/\left(df_{\gamma}-df_{\gamma_f}\right)}{RSS_{\gamma_f}/df_{\gamma_f}}>F_{\left(df_{\gamma}-df_{\gamma_f}\right),df_{\gamma_f}}(\alpha), \tag{2.2}$$

where RSS_{γ} and df_{γ} denote the usual residual sum of squares from fitting γ and the associated degrees of freedom and $F_{\nu_1,\nu_2}(\alpha)$ is the upper α quantile of the F-distribution with ν_1,ν_2 degrees of freedom.

Considering all the subset models from the p predictors as the candidates models, the variable selection confidence set $\widehat{\Gamma}$ is taken to be the set of all those models that satisfy $\widehat{F}(\gamma_f,\gamma)\leq F_{\left(df_{\gamma}-df_{\gamma_f}\right),df_{\gamma_f}}(\alpha)$. By default, the full model is included in $\widehat{\Gamma}$.

Theorem 1. *Under the normal model, if the true model is not the full model, we have $P(\gamma^*\in\widehat{\Gamma})=1-\alpha$. When the true model is the full model, $P(\gamma^*\in\widehat{\Gamma})=1$.*

The result follows trivially from the fact that when $\gamma=\gamma^*$, the F-statistic has a $F_{\left(df_{\gamma^*}-df_{\gamma_f}\right),df_{\gamma^*}}$ distribution. We call $\widehat{\Gamma}$ the exact confidence set (ECS).

The confidence set can be used to check if a given model (e.g., from a selection rule) is too parsimonious. A model in $\widehat{\Gamma}$ is said to be $(1-\alpha)$-SAFE (surviving against F-test evaluation). If a model is not $(1-\alpha)$-SAFE, it most likely misses important predictors. As will be seen, models selected by some popular sparse model selection methods sometimes are not $(1-\alpha)$-SAFE.

The simple VSCS has exact $1-\alpha$ coverage probability, but its size needs to be discussed. The largeness of VSCS is necessary in general. Without any condition on the magnitudes of the effects of the predictors, to guarantee the coverage probability, we must include large models because one cannot tell whether two nested models are both correct, or the smaller model is wrong but the extra terms in the larger one are tiny relative to the sample size. Therefore, $\widehat{\Gamma}$ cannot be improved without further conditions on signal strength. From a practical perspective, this VSCS can be too large to be directly useful beyond checking a model suspected of being overly parsimonious.

2.3. ECS after screening or a conservative selection

In various applications, p is larger than n. Methods such as Lasso and Scad can be applied to obtain a sparse model with a relatively small number of predictors. An important issue then is to examine the reliability of the selected model. In this context, VSCS can provide a complementary perspective on which variables and models may be important.

To construct a VSCS when $p > n$, a variable screening method can be used to sift out unimportant variables and reduce the number of predictors for further consideration to be less than the sample size. In the literature, several screening methods have been proposed with theoretical justifications (see, e.g., Fan and Lv (2008) and Fan and Song (2010)).

Consider a variable screening method ψ that yields a reduced collection of the original predictors, denoted by $\Omega(\psi)$, of size at most $n-1$. The size is typically substantially smaller than n, say of a smaller order. For example, in the sure independence screening procedure of Fan and Lv (2008) based on marginal correlations, the prescribed size of $\Omega(\psi)$ is $d_n = O(n/\log(n))$. Treating $\Omega(\psi)$ as the full model, we can find the ECS as described in the previous subsection and denote it by $\hat{\Gamma}_{\Omega(\psi)}$.

Alternatively, we may consider a set of L high-dimensional model selection methods $\Psi = \{\psi_1, \ldots, \psi_L\}$ that each produces a model with a choice of a tuning parameter. For our purpose, the tuning parameter for each method is chosen conservatively so that the selected model is more likely to not miss the true predictors (but may include noise variables at the same time). The set Ψ (with the tuning parameters) is said to be collectively over-consistent if with probability going to 1 the union of the sets of predictors in the selected models by the L methods, denoted by $\Omega(\Psi)$, contains all the predictors in the true model. Clearly, if any of the model selection methods is actually consistent or over-consistent in selection, then Ψ is collectively over-consistent, but the reverse is not true. Hence the condition is much milder than demanding at least one of the methods to be consistent. Let $\hat{\Gamma}_{\Omega(\Psi)}$ denote the ECS based on $\Omega(\Psi)$ as the full model (assumed to be of size less than n).

For the result below, we assume that the screening or pre-selection by Ψ is done based on a side data set (e.g., from a previous study) or using a small part of the present data. In applications, when the sample size is small and there is no side data, variable screening may be done with the full data, as done in Fan and Lv (2008), although there might be a bias due to reuse of the same data for both steps (screening and VSCS construction).

Corollary 1. *If screening method ψ has $\Omega(\psi)$ containing all the variables in γ^* with probability going to 1, or if Ψ is collectively over-consistent, then* $\liminf_{n\to\infty} P(\gamma^* \in \hat{\Gamma}_{\Omega}) \geq 1-\alpha$, *where Ω is $\Omega(\psi)$ or $\Omega(\Psi)$, respectively.*

2.4. Detectability conditions

We look for conditions under which the terms in the true model will eventually not be missed in the models in the ECS. Let γ denote a model that misses at least one true term. The F-statistic $\widehat{F}(\gamma_f, \gamma)$ has a non-central F-distribution $F_{\left(df_\gamma - df_{\gamma_f}\right), df_\gamma, \delta_\gamma}$, where δ_γ is the non-centrality parameter summarizing the overall effect from missing one or more terms. A VSCS method is said to asymptotically detect all the true terms if all the true terms are included in each of the models in the confidence set with probability going to 1. Asymptotic detectability does not address the issue of inclusion of unnecessary terms. Here we let the number of predictors p depend on n, say p_n. We assume that $p_n \leq (1-\varepsilon)n$ for some possibly small $0 < \varepsilon < 1$. Let p_0 denote the number of terms in the true model, assumed to satisfy that $\log p_0$ is of order $\log n$ and $p_0/p_n \to 0$.

Theorem 2. *Let Γ_u denote the set of models that miss at least one true term. For the ECS, a necessary condition for asymptotic detectability of the true terms at each $0 < \alpha < 1$ is*

$$\min_{\gamma \in \Gamma_u} \frac{\delta_\gamma}{\sqrt{df_\gamma - df_{\gamma_f}}} \to \infty \text{ as } n \to \infty.$$

The true terms are asymptotically detectable if, for some postive constant C

$$\min_{\gamma \in \Gamma_u} \frac{\delta_\gamma}{\xi_n + \sqrt{(df_\gamma - df_{\gamma_f})\left(1 + \log p_n/(df_\gamma - df_{\gamma_f})\right)}} > C$$

for some slowly increasing sequence of $\xi_n \to \infty$. There is a setting with $p_n \to \infty$ such that this condition is necessary for the true terms to be asymptotically detectable.

These necessary and sufficient conditions typically hold when the list of models is fixed (and contains the true model), as $n \to \infty$. There is a small gap (at the order of a logarithmic term in p_n) between the sufficient and necessary conditions if $p_n \to \infty$ (although the term ξ_n is technically needed when the other term in the denominator stays bounded, since it is allowed to approach ∞ arbitrarily slow it is ignored in the discussion). As is shown in the proof, we can construct a setting where the extra logarithmic term is necessary. Therefore the sufficient condition in the theorem is not generally improvable.

Consider a proper subset model of the true model with at least one true term missing. Under the usual assumption that predictors are normalized and not highly correlated, the detectability condition implies that $n\beta^2/(\sqrt{p_n - p_0}\log(p_n)) \to \infty$, where β is any true coefficient. If the true model is sparse and p_n is only of a slightly larger order than p_0, this condition is mild. If the full model is so large that p is of order n, even if the true model is sparse, the true coefficients have to be much larger to ensure the detectability of all the true terms. One should avoid a large full model, if possible. This understanding can be exploited to derive empirical rules to construct a "full model" based on a high-dimensional model selection method so as to gain more insight than offered by the selected model alone (see Section 4.2 for an example).

3. The Subset of Lower Boundary Models (LBMs)

Let γ be a model in a confidence set $\widehat{\Gamma}$. We say that γ is a lower boundary model if there is no model in $\widehat{\Gamma}$ that is nested within γ. Let $LBM(\widehat{\Gamma})$ denote the set of all lower boundary models. From Theorem 1, with probability at least $1-\alpha$, the true model is a LBM or it contains at least one LBM (as its subset model). Therefore, the set of the lower boundary models can naturally serve as a tool to check if a selected model is over-simplifying: If it is not on the lower boundary or above, we can confidently say that the model has missed important predictors and we have an idea of what they are. For the purpose of model identification beyond predictive performance, such an objective check can be helpful to avoid consequences of a decision based on an excessively simplified description of the data.

When the true model is weak (relative to the sample size and the error variance), $LBM(\widehat{\Gamma})$ can involve noise variables. When the signal is strong, however, we have this result.

Corollary 2. *Assume that the ECS asymptotically detects all the true terms. As $n \to \infty$, if γ^* is not the full model, then $P(LBM(\widehat{\Gamma}) = \{\gamma^*\}) \to 1-\alpha$; if γ^* is the full model, then $P(LBM(\widehat{\Gamma}) = \{\gamma^*\}) \to 1$.*

Thus, for a large sample size, when the true terms are asymptotically detectable, the true model will be the only LBM at the given confidence level, and all the useful variables will not be missed in the LBMs with probability close to one.

When constructing a VSCS, it is natural to require that if a model is included in $\widehat{\Gamma}$, then any larger model is also included. We call such a confidence set expansive, the ECS in the previous subsection is not necessarily expansive.

For an expansive confidence set, all we need to know is the set of lower boundary models. The characteristics of the LBMs can be informative regarding the roles of the predictors; we discuss some possibilities.

1. $LBM(\widehat{\Gamma})$ is unique. Here all the predictors in the model are important and no other predictor is proven to be necessary with the limited information available. When the sample size gets much larger, $LBM(\widehat{\Gamma})$ may involve more predictors.

2. The size of $LBM(\widehat{\Gamma})$ is larger than 1, but small. One possibility is that the models in $LBM(\widehat{\Gamma})$ differ in only one or two predictors, in which case the common predictors in the LBMs are important and several predictors are useful but we do not know which one is the best. Another possibility is that the LBMs are quite different in terms of variable composition, which indicates that various combinations of the predictors can give similar explanation power of the response variable.
3. The size of $LBM(\widehat{\Gamma})$ is moderate. This can happen when the number of predictors is not small and a number of predictors are moderately or highly correlated.
4. The size of $LBM(\widehat{\Gamma})$ is relatively large. For high-dimensional cases, this may be typical and one cannot realistically find the "true" or best model; any model selection rule is picking out a model from among many possibilities that have similar criterion values.

3.1. A multiple-explanation index and inclusion importance

Based on the LBMs, we propose some quantities that can be useful for measuring the degree to which multiple models seem to explain the data well, and also the importance of a variable. For a set A, $|A|$ denotes the size of the set. For a given predictor x_i, let $K(x_i)$ be the number of times that x_i appears in the models in $LBM(\widehat{\Gamma})$.

Definition 1. The $(1-\alpha)$-multiple-explanation index (MEI) is
$MEI = \log |LBM(\widehat{\Gamma})|$.

The MEI can be as large as the logarithm of the combinatorial number of p choose $\lfloor p/2 \rfloor$, roughly $(p/2)\log(2e)$. The MEI describes (on a log-scale) how many most-parsimonious models there are to explain the data at the given confidence.

Definition 2. The $(1-\alpha)$-inclusion importance of a predictor x_j is
$II(x_j) = K(x_j)/|LBM(\widehat{\Gamma})|$.

A predictor that appears in all models in $LBM(\widehat{\Gamma})$ has $II = 1$; for others it is less. A variable with $II = 0$ should not be declared useless, only that there is not enough evidence to support that it is useful at the time being.

Corollary 3. *Assume that the ECS asymptotically detects all true terms. Then we have* $\liminf_{n\to\infty} P(MEI = 0) \geq 1-\alpha$ *and* $\lim_{n\to\infty} P\left(II(x_j) = 1\right) = 1$ *for all* x_j *in the true model and* $\lim_{n\to\infty} P\left(II(x_j) > 0\right) \leq \alpha$ *for all* x_j *not in the true model.*

We also consider inclusion importance based on the entire confidence set $\widehat{\Gamma}$, defined by $\widetilde{II} = \widetilde{K}(x_j)/|\widehat{\Gamma}|$, $j = 1, \ldots, p$, where the function $\widetilde{K}(x_j)$ is the number of times that x_j appears in the models of $\widehat{\Gamma}$. Here, unimportant predictors tend to have $\widetilde{II}$ close to 1/2 because, when expanding from the lower boundary models, given the other added predictors the predictor being examined may or may not be included. See the examples in Sections 4 and 5.

3.2. Importance profile and co-importance of predictors

Let $\widehat{\Gamma}_\alpha$ and $LBM(\widehat{\Gamma}_\alpha)$ denote a $1-\alpha$ confidence set ECS, $\widehat{\Gamma}$, and its corresponding lower boundary set, $LBM(\widehat{\Gamma})$. Tracing the LBMs as α changes between these two extremes can be informative.

We introduce two graphical tools to study the explanatory role of predictors. The first is the predictor II profile plot, which traces the inclusion importance, $II(x_j)$ or $\widetilde{II}(x_j)$ of all (or some) predictors, against α. Here one can inspect whether one or more predictors become more important as the confidence level changes. Thus a sharp and steady increase in II when the confidence level changes from 99.9% to 95% suggests that the predictor is highly relevant and should not be missed (see Figures 1(a) and 3(a)).

The second tool is the co-inclusion importance (CII) plot. The CII plot displays the co-importance of variable pairs $\{x_j, x_k\}$, $j, k = 1, \ldots, p$. Let $K(x_j, x_k)$ denote the number of models in the $LBM(\widehat{\Gamma}_\alpha)$ including both x_j and x_k. The co-importance of x_j and x_k is taken as

$$CII(x_j, x_k) = \frac{K(x_j, x_k)}{K(x_j) + K(x_k) - K(x_j, x_k)} \tag{3.1}$$

if $K(x_j, x_k) > 0$, and $CII(x_j, x_k) = 0$ if $K(x_j, x_k) = 0$. Here the denominator counts all the models in $LBM(\widehat{\Gamma}_\alpha)$ that include either x_j or x_k, so $0 \leq CII(x_j, x_k) \leq 1$. For the example of genetic data in Section 4.2, we display co-inclusion importance using a display in which the nodes represent variables and the thickness of the edges is proportional to CII values (see Figures 1(b) and 3 (b)).

Model selection methods often exclude predictors that are highly correlated with ones that are already in a model, whether or not they should be included from a different angle. Some methods, such as Elastic Net(Zou and Hastie (2005)), have been proposed to alleviate the problem. The examination of the LBMs can offer insight on the question of whether two predictors should co-appear or not. The idea also works for a set of three predictors or more.

Table 1. Exact confidence sets (ECSs) for the prostate cancer data. The columns represent confidence level ($(1-\alpha)\%$), size of $\widehat{\Gamma}$ (ECS size), multiple explanation index (MEI), and relative frequency of the predictors in the ECS (columns 4–11).

$(1-\alpha)\%$	ECS size	MEI	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45
99.9	86	1.10	1.00	0.74	0.48	0.56	0.63	0.49	0.48	0.49
99.0	53	0.69	1.00	0.81	0.49	0.53	0.79	0.43	0.40	0.53
95.0	32	0.00	1.00	1.00	0.50	0.50	1.00	0.50	0.50	0.50

4. Data Examples

4.1. Prostate cancer data

Consider the benchmark data set from a study of prostate cancer studied in the model selection literitures by Stamey et al. (1989). Tibshirani (1996), Zou and Hastie (2005) and Li and Lin (2010), among others. The predictors are the clinical measures log(cancer volume) (lcavol), log(prostate weight) (lweight), age, the logarithm of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log(capsular penetration) (lcp), Gleason score (gleason) and percentage Gleason score 4 or 5 (pgg45). The response is the logarithm of prostate-specific antigen (lpsa). In Table 1, we show size and relative frequency of the predictors for the ECSs at the 95, 99 and 99.9% confidence levels. The size of the ECS is clearly monotone in α. At the 99 and 99.9% levels, lcavol, lweight, lbph and svi appear in more than half of the sets, yielding only 2 and 3 LBMs respectively. At the 95% confidence level, lcavol, lweight and svi appear in all the models in the ECS and there is a single LBM containing these predictors.

In Table 2, we show the lower boundary models for the 95 and 99% confidence levels, and models selected using AIC and BIC, Lasso and Scad (Fan and Li (2001)). To compute Lasso and Scad we used the R package `ncvreg` (available at http://cran.r-project.org). The tuning parameters for Lasso and Scad were chosen by 10-fold cross validation. All selection procedures considered turned out to be SAFE at the 95% confidence level, since the selected models were found in the exact confidence set. Our II statistic shows four variables appearing at least once in the lower boundary at the 99% confidence level (lcavol, lweight, lbph, and svi), suggesting that such variables are indeed relevant. At the 95% confidence level, only lcavol, lweight and svi are relevant for a parsimonious story of the underlying process. The other selection methods tend to agree on the importance of these variables.

In Figure 1(a), we show the inclusion importance profiles for the variables computed at the 95% level. Besides the models in the lower boundary, the ECS also includes all the models obtained by expanding from the lower boundary models; therefore, for the ECS we have $\widetilde{II}$=1/2 for α sufficiently close to 0. As

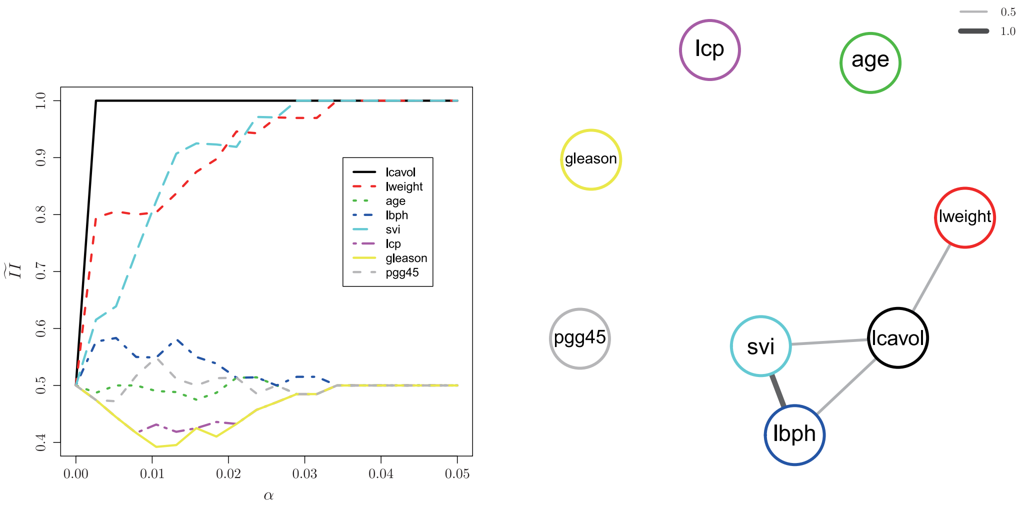
Table 2. Lower boundary models and model selection for the prostate cancer data. We list the lower boundary models (LBMs) (1=predictor included, 0=predictor not included) computed for α=0.01, 0.05, and variable inclusion importance (II) for each predictor. The last columns show the models selected using AIC, BIC, Lasso and Scad (1= predictor included, 0= predictor not included); For the AIC and BIC we used exhaustive search for all possible models; for Lasso and Scad we used 10-fold cross validated tuning parameters.

	$\alpha = 0.01$			$\alpha = 0.05$					
Term	LBMs		II	LBMs	II	AIC	BIC	Lasso	Scad
lcavol	1	1	1.00	1	1.00	1	1	1	1
lweight	1	0	0.50	1	1.00	1	1	1	1
age	0	0	0.00	0	0.00	1	0	0	1
lbph	0	1	0.50	0	0.00	1	0	1	1
svi	0	1	0.50	1	1.00	1	1	1	1
lcp	0	0	0.00	0	0.00	0	0	0	0
gleason	0	0	0.00	0	0.00	0	0	0	0
pgg45	0	0	0.00	0	0.00	0	0	1	1

α increases, we observe different behaviors of the predictors. The importance of lcavol, lweight, and svi increases rapidly as α grows reaching the limit value $\widetilde{II}$=1 for α larger than 0.035. In contrast, the importance of age, lbph, lcp, gleason, and pgg45 converges to 0.5, meaning that we have insufficient information to declare such variables important when α gets larger than 0.035. When α is between 0 and 0.035, lbph and pgg45 appear moderately relevant. In Figure 1(b), we show the co-inclusion importance graph for the variables computed at the 99% level. The nodes correspond to individual variables, while the thickness of the edges is proportional to the co-inclusion importance statistic, CII, defined in Section 3.2. The graph emphasizes pairwise occurrence of variables lcavol, lweight, svi and lbph in the lower boundary. It suggests that, at the 99% level, in terms of explaining the variability in the prostate-specific antigen most parsimoniously, svi and lbph appear together, and they serve as an alternative to lweight.

4.2. Bardet-Biedl syndrome genetic data

We applied our methods to gene expression data from the micro-array experiments of mammalian eye tissue of 120 twelve-week-old male rats (Scheetz et al. (2006)). The outcome of interest is the expression of TRIM32, a gene which has been shown to cause Bardet-Biedl syndrome (Chiang et al. (2006)), a genetic disease of multiple organ systems, including the retina. The micro-arrays contain over 31,042 different probe sets. For each probe, gene expression is measured on a logarithmic scale. Following the pre-processing steps in Scheetz et al. (2006) and Huang and Zhang (2008), we selected 18,976 of the 31,042 probe sets on the



(a) (b)

Figure 1. Inclusion and co-inclusion importance for the prostate cancer data: (a) Inclusion importance ($\widetilde{II}$) for individual variables based on the exact confidence set, $\widehat{\Gamma}$, for α ranging from 0 to 0.05. (b) Co-inclusion importance graph at the 99% confidence level with edges representing values the co-inclusion statstic CII defined in Section 3.2.

array as they "exhibited sufficient signal for reliable analysis and at least 2-fold variation in expression"; then we restricted our attention to the 3,000 probes with the largest variance.

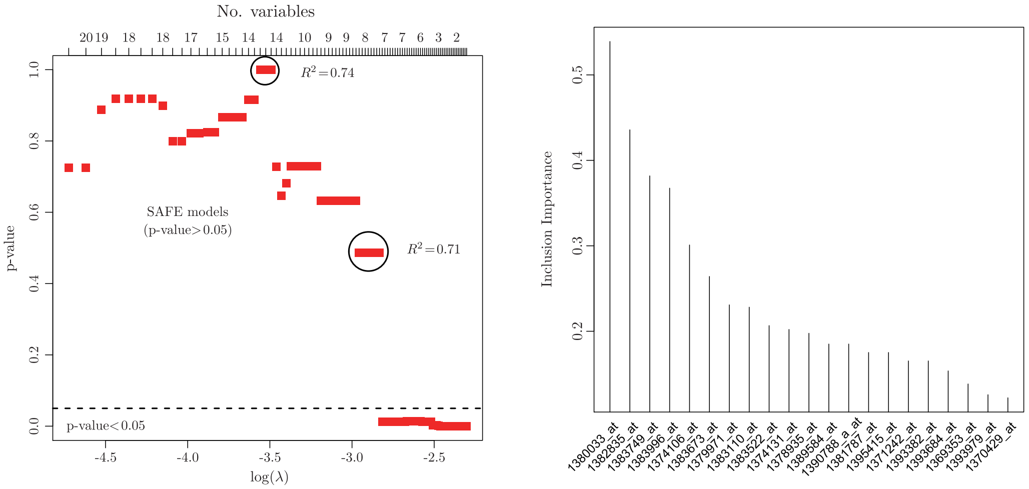
Example 1 (Marginal correlation and Lasso screening)**.** Here we consider statistical screening, which is routinely applied on micro-array data when no biological hypothesis is available. Following Huang and Zhang (2008), we selected 200 variables with the strongest correlation with TRIM32; then we used penalized regression to select a smaller subset of predictors. As an illustration, we considered the Lasso method to carry out the latter step using the R package `ncvreg`. We computed models $\Gamma = \{\gamma_1, \ldots, \gamma_{100}\}$ along the Lasso solution path corresponding to a grid of 100 Lasso regularization parameters and selected the best model $\widehat{\gamma}^* \in \Gamma$ using 5-fold cross-validation consisting of 18 predictors. We then built the "full model" γ_f by moving along the lasso path and taking the largest model on the path with the number of predictors $\tilde{p}$ such that

$$\frac{\widehat{\delta}_{\max}}{\sqrt{(\tilde{p} - p^* + 1)(1 + \log \tilde{p} - \log(\tilde{p} - p^* + 1))}} > C, \tag{4.1}$$

where $\widehat{\delta} = \max_j \{\hat{t}_j^2\}$ is an estimated upper-bound for the non-centrality parameter when one term is missing, $\hat{t}_j$ ($j = 1, \ldots, p^*$) are the t-statistics for the individual variables in the lasso model, and C is a constant representing the necessary signal-to-noise ratio to detect the true terms in the sense of Theorem 2. The left hand side in (4.1) represents an approximated upper bound to the detectability condition in Theorem 2; for example, with $C = 3$, we end up with a full model with 21 predictors. The rationale for the above choice of full model is that if we are to trust the lasso model at all, using a larger full model than given above may even make the strongest term in the lasso model undetectable.

In Figure 2(a), we show the p-values corresponding to the F-test comparisons between the full model and the candidate sub-models along the Lasso path. The upper region in the plot contains 95%-SAFE models along the Lasso path, while the bottom part of the plot contains models that are unsafe due to the overly aggressive Lasso selections, which miss one or more important variables. Of particular interest are the model closest to the boundary (circled in Figure 2(a)), since they represents the most parsimonious Lasso path model within the ECS. Such models include the probes 1370429_at, 1374106_at, 1379971_at, 1383110_at, 1383673_at, 1383996_at, and 1389584_at, which alone explain approximately 70% of the variability in TRIM32. The probes selected by such parsimonious models overlap with the selection obtained by the adaptive Lasso and adaptive Scad methods described Huang and Zhang (2008). The best fitting ECS model on the Lasso path (also circled in Figure 2(a)) accounts for 74% of the variability in TRIM32, but contains almost twice as many variables. In Figure 2(b) we show the inclusion importance of predictors at the 95% confidence level. It shows that, except for two predictors that appear on at least 40% of the LBMs, the other predictors have rather low II values, which reflects the fact that there are many roughly equally plausible models with different compositions of the predictors that can explain TRIM32. From the plot, we may need to admit that the task of identifying the best model at the current sample size is infeasible.

Example 2 (Biological screening)**.** In this example, we consider as potential predictors expression in 11 probes with significant linkage to the known retinal disease genes Bbs1, Bbs4, Bbs8, Opn1sw, Pcdh15, Pde6a, Pex1, Pex7, Rdh12, and Rdp4 in Scheetz et al. (2006) (probe ids 1384603_at, 1383417_at, 1383007_at, 1378416_at, 1388025_at, 1378408_at, 1393426_at, 1376595_at, 1379784_at, 1382949_at, 1371762_at). Figure 3(a) shows the marginal inclusion importance profile plot for α ranging from 0 to 0.1. The most important genes appear to be Bbs8, Bbs4, Pex7, and Opn1sw for all considered confidence levels. We remark that Bbs4 and Bbs8 are known to be related to the Bardet-Biedl syndrome, since they belong to the so-called BBS group. Also Opn1sw is reputed to be important since it represents a non-contiguously regulated gene encoding proteins related to the disease. Figure 3(b) shows the co-inclusion importance graph, where the



(a) (b)

Figure 2. SAFE models on the Lasso path: (a) p-values based on the F-test of the full model γ_f with 21 predictors against smaller models on the Lasso path; the models above the line are safe at the 95% confidence level. (b) Inclusion importance of predictors (II) computed from the lower boundary models.

thickness of the edges represents values of the co-inclusion importance statistic CII defined in Section 3.2. (edges corresponding to $CII \leq 0.2$ are omitted for clarity). It offers information unavailable in the marginal $\widetilde{II}$ plot or from usual model selection processes. The totally isolated predictors in the graph are weak on their own and also do not appear to have any potential jointly with another predictor. There is some evidence to support Rdh12, Opn1sw, and Pex1 as useful, which is already seen from Figure 3(a). But Figure 3(b) further shows that Rdh12 and Pex1 tend to influence the response by appearing together, but they are not connected with Opn1sw, suggesting that {Rdh12, Pex1} and Opn1sw have competing (rather than synergistic) effects. While Bbs4 and Opn1sw have the same high II value, 0.62, their CII value is very small (the connection is very light), which says that their effects are redundant if appearing together. Such information may help the biologist gain more insight on the problem.

In Table 3, we show lower boundary models at the 95% confidence (models LBM1–LBM8), and inclusion importance statistics. For comparison purposes, we also report 5-fold cross-validated Lasso, Scad, and Mcp models selected. Due to pronounced noise in the data, Lasso, Scad, and Mcp generate quite different models for different cross-validation runs; for illustration purposes, we show a single instance. We also show the AIC and BIC models computed by exhaustive search. For each model we report p-values from the F-test defined in Section 2.2

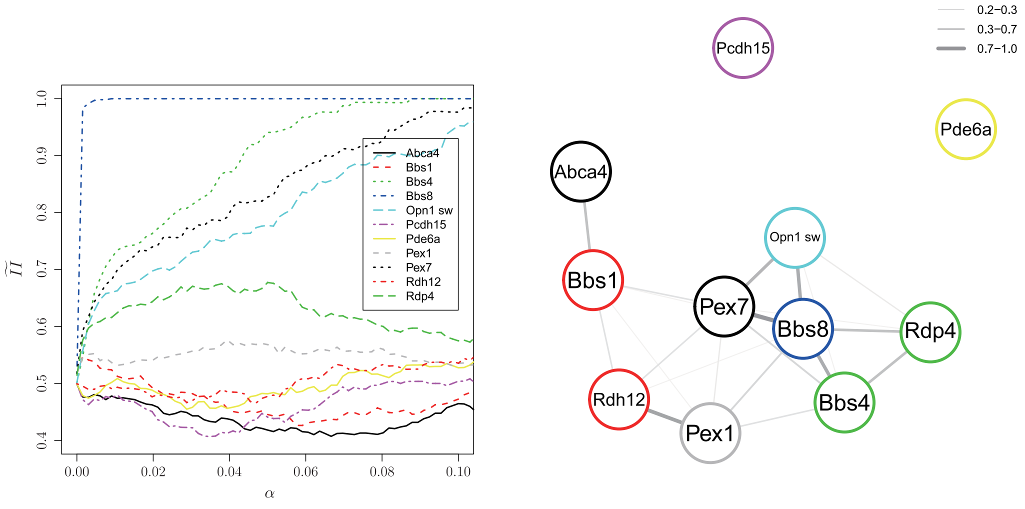
Table 3. Model selection for the Bardet-Biedl data at the 95% confidence level: Lower boundary models (LBM1–LBM8), inclusion importance statistics for the entire confidence set ($\widetilde{II}$) and the lower boundary (II), full, Lasso, Scad, Mcp, AIC, and BIC selections; X denotes "selected" and (*) indicates models outside the confidence set (unSAFE). For each model we include percent p-values for the F-test with the full model (p-val) and coefficients of determination based on ordinary least squares fits (R^2). Lasso, Scad, and Mcp are computed using 5-fold cross-validated hyper-parameters. AIC and BIC models are computed by exhaustive search.

	Abca4	Bbs1	Bbs4	Bbs8	Opn1 sw	Pcdh15	Pde6a	Pex1	Pex7	Rdh12	Rdp4	p-val(%)	R^2
LBM1	X	X		X	X				X			5.38	0.49
LBM2			X	X	X				X			9.37	0.49
LBM3		X	X	X				X	X	X		6.72	0.49
LBM4				X	X			X	X	X		5.15	0.48
LBM5			X	X	X						X	6.96	0.49
LBM6			X	X				X			X	5.82	0.49
LBM7			X	X					X		X	5.51	0.50
LBM8				X	X				X		X	6.96	0.50
II	0.12	0.25	0.62	1.00	0.62	0.00	0.00	0.37	0.75	0.25	0.50		
$\widetilde{II}$	0.42	0.45	0.94	1.00	0.78	0.44	0.48	0.57	0.83	0.48	0.68		
Full	X	X	X	X	X	X	X	X	X	X	X	100.00	0.55
Lasso			X	X	X			X	X		X	23.23	0.51
Scad				X	X				X		X	58.16	0.49
Mcp(*)				X							X	0.42	0.44
AIC			X	X	X		X		X		X	45.30	0.53
BIC(*)				X	X				X			4.70	0.48

and R^2 obtained from a ordinary least square fit. The lower boundary models contain 4 to 6 variables emphasizing various combinations of predictors equally useful in explaining TRIM32. All the LBMs give R^2 near 50%, while the full model with 11 variables yields $R^2 = 55\%$. Genes Bbs4, Bbs8, Opn1sw and Pex7 are included in most LBMs. The same genes also appear frequently in the other selected models. The Mcp and BIC models fall outside the confidence set, so those models cannot be trusted at the 95% confidence level. This is not surprising since BIC and Mcp criteria are known to generating overly sparse selections, so here they are likely missing at least one important variable.

5. Monte Carlo Simulations

We sampled n covariate vectors in the design matrix from a multivariate normal distribution with mean zero and covariance matrix Σ. For each covariate vector, we computed the corresponding response $y = x'\beta + \epsilon$, ϵ sampled from $N(0, \sigma^2)$. We studied the following setups. Model 1: $\beta_j = 1$, $j = 1, \ldots, p/2$ and $\beta_j = 0$, $j = p/2 + 1, \ldots, p$. The correlation between the ith and jth covariates is $\Sigma_{ij} = \rho^{|i-j|}$, $0 \leq \rho < 1$. Model 2: $\beta_j = 1/j$, $j = 1, \ldots, p/2$ and $\beta_j = 0$, $j = p/2 + 1, \ldots, p$. The correlation between the ith and jth covariates is $\Sigma_{ij} = \rho^{|i-j|}$, $0 \leq \rho < 1$. Model 3: Coefficients as in Model 1, but the half of the predictors with zero and the half with nonzero coefficients have $\Sigma_{ij} = \rho \neq 0$. The remaining



(a) (b)

Figure 3. Inclusion and co-inclusion importance at the 95% confidence level for the the Bardet-Biedl micro-array data. (a) Inclusion importance profile of predictors ($\widetilde{II}$) computed on the entire confidence set. (b) 95% confidence co-inclusion importance graph with edges representing values the co-inclusion statistic CII defined in Section 3.2 (edges corresponding to $CII < 0.2$ are omitted for clarity).

pairwise correlations are zero.

5.1. MC Example 1: ECS and LBM set size

In Table 4, we show Monte Carlo estimates for the ECS size, LBM set size, and average number of variables for the lower boundary models based on different choices of p, ρ, and α. The number of models in the ECS is monotone in α with smaller values of α corresponding to larger confidence sets. A similar behavior occurs for the LBM set size when the predictors are orthogonal and all the non-zero coefficients have the same size (Model 1). However, when some of the coefficients are small relative to the others (Model 2), we do not have monotonicicty in α.

While the size of the ECS increases rapidly in p, that of the LBM set remains relatively small. This is important in light of Corollary 2, since the boundary models contain sufficient information about the variables in the true model. In the worst case, in terms of signal-to-noise ratio (Model 2, $p = 12$, $\rho = 0.7$), the boundary set has less than 12 models. This suggests that although the size of all the models in the ECS may be huge without further restrictions when p is large, computing the LBMs can still be managed for a moderately large number

Table 4. Monte Carlo estimates for size of the exact confidence set (ECS), lower boundary model (LBM) set, and the average size of LBMs. The results are based on 500 Monte Carlo samples of size $n = 100$ from Models 1 and 2, for different choices of confidence levels (Conf.), predictors' correlation (ρ), and numbers of predictors (p). Monte Carlo standard errors are smaller than 0.01.

			Model 1				Model 2			
		$p =$	8		12		8		12	
	Conf. (%)	$\rho =$	0.0	0.7	0.0	0.7	0.0	0.7	0.0	0.7
	99.9		17.14	49.36	76.19	392.33	100.27	131.27	1714.08	2176.32
ECS	99.0		15.99	30.89	65.00	211.16	76.03	105.73	1281.22	1737.52
size	95.0		15.29	21.53	60.75	122.81	53.78	81.42	883.06	1327.67
	90.0		14.47	18.02	57.31	93.22	42.91	67.94	687.12	1102.03
	99.9		1.13	3.77	1.49	8.89	1.75	3.51	3.24	9.26
LBM set	99.0		1.03	2.53	1.14	5.84	1.85	3.69	3.89	10.45
size	95.0		1.04	1.82	1.14	3.80	1.86	3.68	4.29	11.11
	90.0		1.07	1.63	1.19	3.07	1.85	3.63	4.44	11.17
	99.9		3.92	3.28	5.84	5.04	1.59	1.71	1.73	2.32
LBMs	99.0		3.99	3.55	5.99	5.42	2.02	2.07	2.29	2.77
av. size	95.0		4.06	3.81	6.09	5.77	2.54	2.43	2.92	3.23
	90.0		4.12	3.99	6.17	5.96	2.83	2.67	3.29	3.47

of predictors.

The number of predictors in the LBMs grows with α. When the confidence level increases, the LBMs are more parsimonious. If the confidence level is small, there are only a few, relatively large LBMs. In the presence of relatively small coefficients and a large correlation, the LBMs are numerous but they contain fewer predictors.

5.2. MC Example 2: importance profile of predictors

We illustrate the behavior of the ECS and LBM sets given different significance levels. We considered a sequence of equally spaced values for α ranging from 0.001 to 0.1, and drew 100 Monte Carlo samples of size $n = 100$ from Model 3 with $p = 8$ predictors. Only the first four predictors had nonzero coefficients $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$. While x_1 x_2, x_5 and x_6 were orthogonal to all the other predictors, x_3, x_4,x_7 and x_8 were moderately correlated with correlation $\rho = 0.5$.

Figure 4(a) shows the Monte Carlo averages of the predictors' inclusion importance profile plots (solid line) for the ECS with 95% confidence bands (dashed lines). The lighter lines show individual Monte Carlo realizations of the profile plots. All the predictors with nonzero coefficients show inclusion importance close to 1, meaning that the relevant predictors are almost always included in the ECS. In contrast, the independent predictors show smaller variability compared to the correlated predictors.

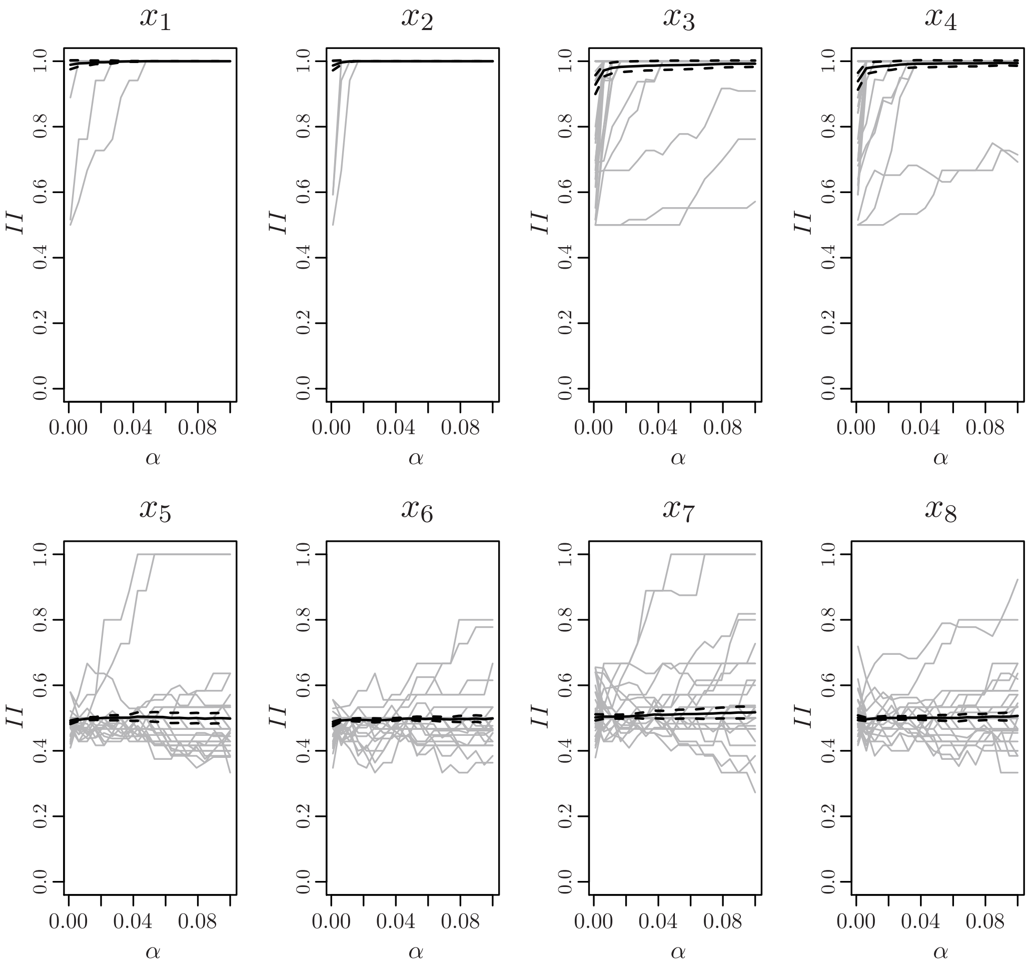


Figure 4. Inclusion importance (II) profile plots simulated from Model 3. The solid and dashed black lines represent Monte Carlo means and 95% confidence bands, respectively. The lighter lines show 100 Monte Carlo realizations. The plots are based on 50 Monte Carlo samples of size $n = 100$ from Model 3, where predictors x_1, x_2, x_5, x_6 are orthogonal to all the predictors, while x_3, x_4, x_7, x_8 have pairwise correlation $\rho = 0.5$.

6. Comparison with the Work of Hansen, Lunde and Nason (2011)

A recent paper by Hansen, Lunde and Nason (2011) is closely related to our work. We share the goal of providing a confidence set of models to give the data analyst a proper sense on how far the information in the data can allow him/her to go in terms of identification of the best model. They also share the view that the size of the confidence set is a valuable indicator of the degree of model selection uncertainty. Here we discuss the differences.

As to assumptions in Hansen, Lunde and Nason (2011) the true model is not specified, and candidate models provide estimations/predictions that are assessed in terms of a chosen loss function. This flexible framework targets various applications. Our work starts with a clearly specified full model in the normal regression setting and the issue is on which subset model to use. Given our setup in this paper, assumptions are not needed for construction of the variable selection confidence sets, whrer Hansen, Lunde and Nason (2011) assume that the mean of the loss difference between any two models stays the same over time. This last seems restrictive, for the usual regression data, conditional on the design matrix, even for the true model the losses at the observations typically are distinct and the mean loss differences are expected to depend on the cases. Further, their demand that for each pair of models the mean loss difference is either zero or a nonzero constant seems at odds with common applications. The mean loss difference assumption in Hansen, Lunde and Nason (2011) also rules out applications where overfitting models exist, and it may over-simplify the nature of different performances of the candidate models.

Given our framework, the exact confidence set offers a finite-sample coverage guarantee. When the set of lower boundary models is considered, we give only asymptotic results on the containment of the true model and related quantities under an additional condition on the signal strength. Hansen, Lunde and Nason (2011) have an asymptotic result on the behavior of their confidence set. Later, in pursuit of a finite-sample coverage probability, a coherency condition is needed to relate the equivalence test and the elimination rule. Since the coherency is an exact requirement, the asymptotic justification of their bootstrap method does not appear to be sufficient for deriving a non-asymptotic confidence set.

We allow the number of predictors, p to grow with n to capture the challenge in high-dimensional regression. In such a setting, there are possibly many models that are hard to distinguish by any method. Our idea is to use the set of lower boundary models to properly reflect reality. Although it is not explicitly stated in Hansen, Lunde and Nason (2011), the number of models considered there is fixed for the theoretical results. The issue with the number of candidate models being large relative to the sample size seems to be a real challenge to the their methodology.

7. Concluding Remarks

For reasonably complicated high-dimensional data, it is usually unrealistic to expect a unique model to stand out as the “true” or best model. Rather a number of models are more or less equally supported by the data. In such a situation, it is better to be aware of the top models for a deeper understanding of the relationship between the response and the predictors. In this work, we have demonstrated the usefulness of having a variable selection confidence set from multiple aspects. Specifically, the examination on whether a model selected is $(1-\alpha)$-SAFE, the inclusion importance and co-inclusion importance all provide valuable information unavailable in the single selected model, no matter the model selection criterion.

Statistical estimation/prediction or inference based on more than one model is not a new topic. Model averaging tries to reduce the uncertainty associated with the choice of a single model. For example, Burnham and Anderson (2002) advocate the use of Akaike weights for assessing strengths of the candidate models, and for model averaging. While such model selection criterion-based weights provide an intuitive view on the relative usefulnesses of the candidate models, more work is needed to understand how the weights can be interpreted pertaining to reliably selecting the most important variables.

Our VSCS can be used as a model selection diagnostic tool. To examine a model selected by a sparse modeling method, one can first come up with a super model by moving further along its solution path and adding a few predictors recommended by some other model selection methods. If the model is not 95%-SAFE, then there is strong reason to doubt the soundness of the set of predictors in the model. Furthermore, by comparing it with the LBMs, one has a good idea of which important predictors are missed. Of course, the outcome of the diagnostic process is much more informative when a negative result is reached.

Although we can always quickly check whether one or a few models selected by certain methods are in the ECS or not, when p is large it is computationally challenging to go over all the subset models to identify the entire ECS without further conditions. With that, in the numerical work of this paper, we have limited our scope to manageable sets of candidate models with p relatively small (possibly after a variable screening). Although the size of the ECS may grow quickly in p, the number of LBMs does not grow as much and can often be computed, as seen in the illustrations of Section 5. Nevertheless, the computation to list out all the LBMs can still be costly. We plan to seriously examine the computation issues of ECS and LBMs in the future.

Acknowledgement

We sincerely thank the two reviewers and the AE for their very helpful comments and suggestions for improving our work. In particular, the reference of Hansen, Lunde and Nason (2011) that they brought to our attention for comparison and discussion is appreciated. The work of Yuhong Yang was partially supported by the NSF Grant DMS-1106576.

References

Birgé, L. (2001). An alternative point of view on Lepski's method. In *State of the Art in Probability and Statistics*, 113-133. Institute of Mathematical Statistics.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: a Practical Information-theoretic Approach.* Springer-Verlag.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference (Disc: P444-466). *J. Roy. Statist. Soc. Ser. B* **158**, 419-444.

Chiang, A., Beck, J., Yen, H., Tayeh, M., Scheetz, T., Swiderski, R., Nishimura, D., Braun, T., Kim, K., Huang, J. et al. (2006). Homozygosity mapping with snp arrays identfies trim32, an e3 ubiquitin ligase, as a Bardet-Biedl syndrome gene (BBS11). *Proc. Natl. Acad. Sci.* **103**, 6287-6292.

Draper, D. (1995). Assessment and propagation of model uncertainty (Disc: P71-97). *J. Roy. Statist. Soc. Ser. B* **57**, 45-70.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18**, 71-103.

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. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. Ser. B* **70**, 849-911.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38**, 3567-3604.

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

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.* **98**, 879-899.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statist. Sci.* **14**, 382-401.

Huang, J. and Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica* **18**, 1603-1618.

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

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28**, 1302-1338.

Lehmann, E. L. and Romano, J. P. (2006). *Testing Statistical Hypothesis.* Springer.

Li, Q. and Lin, N. (2010). The Bayesian elastic net. *Bayesian Anal.* **5**, 151-170.

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

Scheetz, T., Kim, K., Swiderski, R., Philp, A., Braun, T., Knudtson, K., Dorrance, A., DiBona, G., Huang, J. and Casavant, T. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proc. Natl. Acad. Sci.* **103**, 14429-14434.

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

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha E. F., Redwine, E. and Yang, N (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate ii: radical prostatectomy treated patients. *J. Urology* **141**, 1076-1083.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Yang, Y. (2001). Adaptive regression by mixing. *J. Amer. Statist. Assoc.* **96**, 574-588.

Yuan, Z. and Yang, Y. (2005). Combining linear regression models: When and How? *J. Amer. Statist. Assoc.* **100**, 1202-1214.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67**, 301-320.

Richard Berry Building, University of Melbourne, Parkville, 3010, VIC, Australia.

E-mail: dferrari@unimelb.edu.au

313 Ford Hall, 224 Church Street SE , Minneapolis, MN, 55455, USA.

E-mail: yyang@stat.edu.au

(Received March 2014; accepted November 2014)