

Statistica Sinica Preprint No: SS-2019-0210

Title	Performance Assessment of High-dimensional Variable Identification
Manuscript ID	SS-2019-0210
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0210
Complete List of Authors	Yanjia Yu, Yi Yang and Yuhong Yang
Corresponding Author	Yanjia Yu
E-mail	yuxxx748@umn.edu
Notice: Accepted version subject to English editing.	

Performance Assessment of High-dimensional Variable Identification

Yanjia Yu, Yi Yang and Yuhong Yang

School of Statistics, University of Minnesota, USA

Department of Mathematics and Statistics, McGill University, Canada

School of Statistics, University of Minnesota, USA

Abstract: Since model selection is ubiquitous in data analysis, reproducibility of statistical results demands a serious evaluation of reliability of the employed model selection method, no matter what label it may have in terms of good properties. Instability measures have been proposed for evaluating model selection uncertainty. However, low instability does not necessarily indicate that the selected model is trustworthy, since low instability can also arise when a certain method tends to select an overly parsimonious model. F - and G -measures have become increasingly popular for assessing variable selection performance in theoretical studies and simulation results. However, they are not computable in practice. In this work, we propose an estimation method for F - and G -measures and prove their desirable properties of uniform consistency. This gives the data analyst a valuable tool to compare different variable selection methods based on the data at hand. Extensive simulations are conducted to show the very good finite sample performance of our approach. We further demonstrate the application of our methods using several microarray gene expression data sets, with intriguing findings.

Key words and phrases: F -measure; G -measure; Gene expression; Model averaging;

Reproducibility; Variable selection performance.

1. INTRODUCTION

Variable selection is of interest in many fields such as bioinformatics, genomics, finance and economics, etc. In bioinformatics, for example, microarray gene expression data are collected to identify cancer-related biomarkers in order to differentiate affected patients from healthy individuals based on their gene expression profile. The number of variables, p , in typical microarray gene expression data is of 10^3 – 10^5 magnitude, while the number of subjects, n , is of 10^1 – 10^3 magnitude. For such problems where $p \gg n$, the penalized likelihood estimation provides a class of methods for selecting the variables (see e.g., Fan and Lv, 2010). However, it is well recognized in the literature that model selection methods, including penalization methods for high-dimensional data, often encounter instability issues (Chatfield, 1995; Draper, 1995; Breiman, 1996a,b; Buckland et al., 1997; Yuan and Yang, 2005; Lim and Yu, 2016). For example, removing a few observations or adding small perturbations to the data may result in dramatically different sets of variables being selected (Meinshausen and Bühlmann, 2006; Nan and Yang, 2014; Lim and Yu, 2016). This uncertainty in variable selection, as is well known, may have severe practical consequences in applications. At a larger scale, the reproducibility crisis is a major problem in the science community (McNutt, 2014; Stodden, 2015).

Variable selection uncertainty is mainly evaluated by instability measures in existing literature, which test how sensitive a variable selection method is to small changes of the data, either by subsampling (Chen et al., 2007), resampling (Breiman, 1996b; Buckland et al., 1997) or perturbations (Breiman, 1996b). However, a low instability measure does not necessarily indicate that a variable selection result is reliable, since low instability can also arise when a method tends to select an overly parsimonious model (e.g., the intercept-only model in the extreme case).

There is therefore a great need for measures that can fully evaluate the uncertainty of variable selection beyond instability. In variable selection, one mainly cares about two types of errors: including unnecessary variables and excluding important ones. F - and G -measures, appearing often in the field of information retrieval (Billsus and Pazzani, 1998), are becoming popular tools for assessing the overall variable selection performance (see e.g., Lim, 2011; Lim and Yu, 2016). Specifically, the F -measure is the harmonic mean of *precision* and *recall*, where precision (or positive predictive value) is defined as the fraction of selected variables that are true variables, and recall (also known as sensitivity) is defined as the fraction of the true variables that are selected. The G -measure is the geometric mean of precision and recall. By combining precision and recall into one measure, one can evaluate the overall accuracy of a given variable selection method. Clearly, a higher F (or G) value indicates better selection

performance in an overall sense. However, previous work in the literature calculates the F - (or G -)measure of a given selection method for simulated data alone (where the true model is known) and they do not work for real data.

In this paper, we propose a method for performance assessment of (high-dimensional) variable identification (PAVI) by estimating the F - or G -measure based on the combination of multiple candidate models under a proper weighting scheme. Our proposal works for both regression and classification and applies to both synthetic and real data. Under sensible conditions, we show that our estimates are uniformly consistent in estimating the true F - and G -measures for any set of models to be checked. The candidate models can be very flexible. For example, they can be obtained by penalization using Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), adaptive Lasso (Zou, 2006), MCP (Zhang, 2010) or by other variable selection techniques. Two weighting schemes are considered in this work: the adaptive regression by mixing (Yang, 2001) and weighting via information criteria (see e.g., Nan and Yang, 2014). In the simulation section, we show a very reliable estimation performance of our method for both classification and regression. We further demonstrate our methods by analyzing several microarray gene expression data from real applications. The real data analysis suggests that PAVI is a very useful tool for evaluating the variable selection performance of high-dimensional linear-based models. They provide useful information on the reliability and reproducibility of a given model when the true

model is unknown. For example, one may justifiably doubt the reproducibility of a model that has very small estimated F and G values.

The remainder of the paper is organized as follows. In Section 2, we define the F - and G -measures and introduce our estimation methods. Section 3 provides the theoretical justification for the PAVI estimators of the F - and G -measures. Section 4 gives some implementation details of PAVI for both regression and classification, including how to obtain the candidate models and assign weights. Simulation results are presented in Section 5. We demonstrate our methods by analyzing three well-studied gene expression datasets in Section 6. Conclusions are given in Section 7. The technical proofs are relegated to the supplement, where additional numerical results are presented as well.

2. METHODOLOGY

Let us consider the generalized linear model framework. Denote $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ the $n \times p$ design matrix with $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$, $i = 1, \dots, n$. Let $\mathbf{y} = (y_1, \dots, y_n)^\top$ be the n -dimensional response vector. For regression with a continuous response, we consider the linear regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is the vector of n independent errors and $\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_p^*)^\top$ is a p -dimensional coefficient vector of the true underlying model that generates the

data. For classification, we consider the binary logistic regression model for ease of presentation. Let $Y \in \{0, 1\}$ be a binary response variable and $X \in \mathbb{R}^p$ be a p -dimensional predictor vector. We assume that Y follows the Bernoulli distribution given $X = \mathbf{x}$, with conditional probability

$$\Pr(Y = 1|X = \mathbf{x}) = 1 - \Pr(Y = 0|X = \mathbf{x}) = \frac{e^{\mathbf{x}^\top \boldsymbol{\beta}^*}}{1 + e^{\mathbf{x}^\top \boldsymbol{\beta}^*}}. \quad (2.1)$$

Let $\mathcal{A}^* = \text{supp}(\boldsymbol{\beta}^*) \equiv \{j : \beta_j^* \neq 0\}$ be the index set of the variables in the true model with size $|\mathcal{A}^*|$, where $|\cdot|$ denotes the cardinality of a set. For both regression and classification, we assume that the true model is sparse. In other words, most coefficients in $\boldsymbol{\beta}^*$ are exactly zero, so that $|\mathcal{A}^*|$ is small.

Let $\mathcal{A}^0 = \{j : \beta_j^0 \neq 0\}$ be an index set of all nonzero coefficients from any given variable selection result $\boldsymbol{\beta}^0$. One can use F - and G -measures to evaluate the performance of \mathcal{A}^0 . F - and G -measures take values between 0 and 1, and a higher value indicates better performance of the variable selection method. The definitions of F - and G -measures rely on two quantities, *precision* and *recall*. The precision pr for \mathcal{A}^0 is the fraction of true variables in the given model \mathcal{A}^0 , i.e., $pr(\mathcal{A}^0) \equiv pr(\mathcal{A}^0; \mathcal{A}^*) = |\mathcal{A}^0 \cap \mathcal{A}^*|/|\mathcal{A}^0|$, and the recall re for \mathcal{A}^0 is the fraction of variables in the true model \mathcal{A}^* that are selected, i.e., $re(\mathcal{A}^0) \equiv re(\mathcal{A}^0; \mathcal{A}^*) = |\mathcal{A}^0 \cap \mathcal{A}^*|/|\mathcal{A}^*|$. The F -measure for a given model \mathcal{A}^0 is defined as the harmonic mean of precision and recall, while the G -measure is

defined as the geometric mean of the two. Specifically,

$$F(\mathcal{A}^0) = F(\mathcal{A}^0; \mathcal{A}^*) \equiv \frac{2 \times pr(\mathcal{A}^0) \times re(\mathcal{A}^0)}{pr(\mathcal{A}^0) + re(\mathcal{A}^0)} = \frac{2|\mathcal{A}^0 \cap \mathcal{A}^*|}{|\mathcal{A}^0| + |\mathcal{A}^*|},$$

and

$$G(\mathcal{A}^0) = G(\mathcal{A}^0; \mathcal{A}^*) \equiv \sqrt{pr(\mathcal{A}^0) \times re(\mathcal{A}^0)} = \frac{|\mathcal{A}^0 \cap \mathcal{A}^*|}{\sqrt{|\mathcal{A}^0| \cdot |\mathcal{A}^*|}}.$$

In penalized regression, it is well known that when the penalty level is increased, fewer active variables are selected. Therefore, false positives are less likely to happen, while false negatives become more likely. By taking the harmonic (or geometric) mean of precision and recall, the F -measure (or G -measure) integrates both false-positive and false-negative aspects into a single characterization. Given \mathcal{A}^0 , a high F - or G -measure indicates that both false-positive and false-negative rates are low. For example, if $\mathcal{A}^* = (1, 1, 1, 0, 0, 0, 0)$ and $\mathcal{A}_1^0 = (1, 1, 1, 0, 0, 0, 1)$, then $pr(\mathcal{A}_1^0) = 3/4$, $re(\mathcal{A}_1^0) = 1$, $F(\mathcal{A}_1^0) = 6/7$ and $G(\mathcal{A}_1^0) = \sqrt{3}/2$. For the same \mathcal{A}^* , if we consider a worse case $\mathcal{A}_2^0 = (1, 1, 0, 0, 0, 0, 1)$, then $pr(\mathcal{A}_2^0) = 2/3$, $re(\mathcal{A}_2^0) = 2/3$, $F(\mathcal{A}_2^0) = 2/3$ and $G(\mathcal{A}_2^0) = 2/3$. The F - and G -measures are smaller than those in the first case due to the existence of both under-selection and over-selection. In general, F - and G -measures are conservative in the sense that both are more sensitive to under-selection than to over-selection. Specifically, suppose $|\mathcal{A}^*| = m$. If \mathcal{A}_3^0 over-selects one variable, then $|\mathcal{A}_3^0| = m + 1$, $F(\mathcal{A}_3^0) = 2m/(2m + 1)$, and $G(\mathcal{A}_3^0) = \sqrt{m/(m + 1)}$, while if \mathcal{A}_4^0 under-selects one variable, then $|\mathcal{A}_4^0| = m - 1$, $F(\mathcal{A}_4^0) = (2m - 2)/(2m - 1)$,

and $G(\mathcal{A}_4^0) = \sqrt{(m-1)/m}$. One can easily see that $F(\mathcal{A}_3^0) > F(\mathcal{A}_4^0)$ and $G(\mathcal{A}_3^0) > G(\mathcal{A}_4^0)$.

In real applications, the true model \mathcal{A}^* is usually unknown, and thus we cannot directly know $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$ for any given model \mathcal{A}^0 . However, by borrowing information from a group of given models, we can estimate $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$ from the data. Suppose that we have a set of candidate models $\mathbb{S} = \{\mathcal{A}^1, \dots, \mathcal{A}^K\}$, which can be obtained from a preliminary analysis. When the model size p is small, we can use a full collection of all-subset models $\mathbb{S} = \mathbb{C}$, where

$$\mathbb{C} = \{\emptyset, \{1\}, \dots, \{p\}, \{1, 2\}, \{1, 3\}, \dots, \{1, \dots, p\}\}$$

with $1, \dots, p$ represent the indices of the p variables. If p is too large, we can choose \mathbb{S} as a group of models obtained from penalized methods such as Lasso, adaptive Lasso, SCAD and MCP, etc. Define $\mathbf{w} = \{w_1, \dots, w_K\}$ as the corresponding data-driven weights for $\mathbb{S} = \{\mathcal{A}^1, \dots, \mathcal{A}^K\}$, where $w_k \geq 0$ for $k = 1, \dots, K$ and $\sum_{k=1}^K w_k = 1$. In Section 4.1, we further describe how we acquire \mathbb{S} and \mathbf{w} . For now we assume these are already properly acquired. For each \mathcal{A}^k , we define the estimated precision and recall for \mathcal{A}^0 (relative to \mathcal{A}^k) as $pr(\mathcal{A}^0; \mathcal{A}^k) = |\mathcal{A}^0 \cap \mathcal{A}^k|/|\mathcal{A}^0|$ and $re(\mathcal{A}^0; \mathcal{A}^k) = |\mathcal{A}^0 \cap \mathcal{A}^k|/|\mathcal{A}^k|$, and propose the following $\widehat{F}(\mathcal{A}^0)$ by PAVI to estimate $F(\mathcal{A}^0)$

$$\widehat{F}(\mathcal{A}^0) = \sum_{k=1}^K w_k F(\mathcal{A}^0; \mathcal{A}^k) = 2 \sum_{k=1}^K w_k \frac{|\mathcal{A}^0 \cap \mathcal{A}^k|}{|\mathcal{A}^0| + |\mathcal{A}^k|}. \quad (2.2)$$

Similarly, we propose $\widehat{G}(\mathcal{A}^0)$ by PAVI to estimate $G(\mathcal{A}^0)$

$$\widehat{G}(\mathcal{A}^0) = \sum_{k=1}^K w_k G(\mathcal{A}^0; \mathcal{A}^k) = 2 \sum_{k=1}^K w_k \frac{|\mathcal{A}^0 \cap \mathcal{A}^k|}{\sqrt{|\mathcal{A}^0| \cdot |\mathcal{A}^k|}}. \quad (2.3)$$

We define the (sample) standard deviation of $\widehat{F}(\mathcal{A}^0)$ as

$$\text{sd}(\widehat{F}(\mathcal{A}^0)) = \sqrt{\sum_{k=1}^K w_k (F(\mathcal{A}^0; \mathcal{A}^k) - \widehat{F}(\mathcal{A}^0))^2}. \quad (2.4)$$

Similarly, the (sample) standard deviation of $\widehat{G}(\mathcal{A}^0)$ is

$$\text{sd}(\widehat{G}(\mathcal{A}^0)) = \sqrt{\sum_{k=1}^K w_k (G(\mathcal{A}^0; \mathcal{A}^k) - \widehat{G}(\mathcal{A}^0))^2}. \quad (2.5)$$

In (2.2) and (2.3), $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ are estimated using the candidate models $\mathcal{A}^k \in \mathbb{S}$ and weights $w_k \in \mathbf{w}$ for $k = 1, \dots, K$. Intuitively, if higher weights w_k 's are assigned to those \mathcal{A}^k 's that are closer to the true model \mathcal{A}^* , then $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ should be able to better approximate the true values of $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$, respectively. In Section 4.2, we discuss the methods for computing weights \mathbf{w} from the data.

3. THEORY

In this section, we show that the proposed estimators \widehat{F} and \widehat{G} are uniformly consistent for the true F and G , respectively, over the set of all models to be checked. The theory relies on the property, referred to as *weak consistency* (see Definition 1 and Nan and Yang, 2014), of the data-dependent model weights

$\mathbf{w} = \{w_1, \dots, w_K\}$, and the *weak inclusion property* which entails if a model screening process is applied to reduce the model list (Definition 2).

Definition 1 (Weak consistency). The weighting vector $\mathbf{w} = (w_1, \dots, w_K)^\top$ is weakly consistent if

$$\frac{\sum_{k=1}^K w_k \cdot |\mathcal{A}^k \nabla \mathcal{A}^*|}{|\mathcal{A}^*|} \xrightarrow{p} 0 \text{ as } n \rightarrow \infty,$$

where ∇ denotes the symmetric difference between two sets.

Remark 1. The definition basically says that \mathbf{w} is concentrated enough around the true model \mathcal{A}^* so that the weighted deviation $|\mathcal{A}^k \nabla \mathcal{A}^*|$ eventually diminishes relative to the size of the true model. When the true model is allowed to increase in dimension as n increases, including the denominator $|\mathcal{A}^*|$ in the definition makes the condition more likely to be satisfied.

The following theorem shows that under the weak consistency condition, the estimators \hat{F} and \hat{G} are uniformly consistent (the proof is in the supplement).

Theorem 1 (Uniform consistency of \hat{F} and \hat{G}). *Suppose the model weighting \mathbf{w} is weakly consistent. Then \hat{F} and \hat{G} based on PAVI are uniformly consistent in the sense that*

$$\sup_{\mathcal{A}^0 \in \mathcal{C}} |\hat{F}(\mathcal{A}^0) - F(\mathcal{A}^0)| \xrightarrow{p} 0 \text{ as } n \rightarrow \infty;$$

$$\sup_{\mathcal{A}^0 \in \mathcal{C}} |\hat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)| \xrightarrow{p} 0 \text{ as } n \rightarrow \infty.$$

From this theorem we see that if the model weighting mostly focuses on models that are sensibly close to the true model, then our estimated \widehat{F} and \widehat{G} will be close to their respective true values. Clearly, we also have $E|\widehat{F}(\mathcal{A}^0) - F(\mathcal{A}^0)| \rightarrow 0$ and $E|\widehat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)| \rightarrow 0$ uniformly.

Theorem 2 (Uniform convergence of $\text{sd}(\widehat{F})$ and $\text{sd}(\widehat{G})$). *Suppose the model weighting \mathbf{w} is weakly consistent. Then $\text{sd}(\widehat{F})$ and $\text{sd}(\widehat{G})$ based on PAVI converge to 0 in probability uniformly in the sense that*

$$\sup_{\mathcal{A}^0 \in \mathbb{C}} |\text{sd}(\widehat{F}(\mathcal{A}^0))| \xrightarrow{p} 0 \quad \text{as } n \rightarrow \infty;$$

$$\sup_{\mathcal{A}^0 \in \mathbb{C}} |\text{sd}(\widehat{G}(\mathcal{A}^0))| \xrightarrow{p} 0 \quad \text{as } n \rightarrow \infty.$$

From this theorem we see that if the model weighting is sensible, then $\text{sd}(\widehat{F})$ and $\text{sd}(\widehat{G})$ will be close to zero. The results also support reliability of our PAVI method.

Theorems 1 and 2 rely on the weak consistency of \mathbf{w} . Clearly, when the candidate models in \mathbb{S} are all poor, weak consistency may not be plausible. One can choose all-subset models \mathbb{C} as \mathbb{S} when p is small, since it always contains \mathcal{A}^* . However, in the high-dimensional case, it would be computationally infeasible to use \mathbb{C} and a model screening process may be applied (e.g., considering solution paths of model selection methods).

Definition 2 (Weak inclusion property). A set of candidate models \mathbb{S} obtained by a model screening process is called weakly inclusive with respect to \mathbf{w} on \mathbb{C}

if $\sum_{k \in \mathbb{S}} w_k$ is bounded away from zero in probability.

Theorem 3. *Under the assumption that the weighting vector \mathbf{w} on the all-subset models \mathbb{C} is weakly consistent, as long as \mathbb{S} is weakly inclusive, the conclusions of Theorems 1 and 2 still hold.*

Remarks on this result are given in the supplement.

4. IMPLEMENTATION

4.1 Candidate models

We discuss how to choose the candidate models for computing \hat{F} and \hat{G} . To get the candidate models, we can use a complete collection of all-subset models, i.e., choose $\mathbb{S} = \mathbb{C}$. However, in the high-dimensional case where $p \gg n$, it is almost impossible to use all subsets due to high computational cost.

We show in the following how it is done for linear and logistic regression models in the high-dimensional setting. Similar procedures apply to other likelihood-based models. Given n independent observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ for the pair (X, Y) , we can fit the linear or logistic regression model by minimizing the penalized negative log-likelihood

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} -\ell(\boldsymbol{\beta}) + \sum_{j=1}^p p_{\lambda}(\beta_j), \quad (4.6)$$

4.2 Weighting methods

where $-\ell(\boldsymbol{\beta}) = (2n)^{-1} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$ for linear regression and

$$-\ell(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^n \{-y_i \log \pi_i - (1 - y_i) \log(1 - \pi_i)\}$$

for the logistic regression, where $\pi_i = \Pr(Y_i = 1 | X_i = \mathbf{x}_i)$ is the probability in (2.1) for observation i . The nonnegative penalty function $p_\lambda(\cdot)$ with $\lambda \in [0, \infty)$ can be the Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), MCP (Zhang, 2010), or other regularizers.

We compute the models $\mathbb{S} = \{\mathcal{A}^{\lambda_1}, \dots, \mathcal{A}^{\lambda_L}\}$ for, e.g., Lasso, SCAD and MCP, respectively, on the solution paths $\{\hat{\boldsymbol{\beta}}^{\lambda_1}, \dots, \hat{\boldsymbol{\beta}}^{\lambda_L}\}$ for decreasing sequences of tuning parameters $\{\lambda_1, \dots, \lambda_L\}$. These models are then combined together as a set of candidate models $\mathbb{S} = \{\mathbb{S}_{\text{Lasso}}, \mathbb{S}_{\text{SCAD}}, \mathbb{S}_{\text{MCP}}\}$. One can efficiently compute the whole solution paths of Lasso using **glmnet** (Friedman et al., 2010), and of SCAD and MCP using **ncvreg** (Breheny and Huang, 2011).

4.2 Weighting methods

There are several different methods in the literature for determining the weights $\mathbf{w} = \{w_1, \dots, w_K\}$. For example, Buckland et al. (1997) and Leung and Barron (2006) proposed information-criterion-based methods for weighting, such as those using AIC (Akaike, 1973) and BIC (Schwarz, 1978); Hoeting et al. (1999) proposed the Bayesian model averaging (BMA) method for weighting; and Yang (2001) studied a weighting strategy named the adaptive regression by mixing (ARM), which computes the weights by data splitting and cross-assessment. It

4.2 Weighting methods

was proven in Yang (2001) that the weighting by ARM delivers the best rate of convergence for regression estimation. In Yang (2000), the ARM weighting method was also extended to the classification setting. When the number of models in the candidate-model set is fixed, BMA weighting is consistent (thus weakly consistent). From Yang (2007), when one properly chooses the data splitting ratio, the ARM weighting can be consistent. More recently, Lai et al. (2015) proposed Fisher's fiducial-based methods for deriving probability density functions as weights on the set of candidate models. They showed that, under certain conditions, their method is consistent when p is diverging and the size of the true model is either fixed or diverging. In this paper, we only consider the ARM weighting and weighting based on an information criterion.

Weighting using ARM for linear regression

To get the ARM weights, we randomly split the data $\mathbf{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ into a training set \mathbf{D}_1 and a test set \mathbf{D}_2 of (approximately) equal size. We train the linear regression model on \mathbf{D}_1 and evaluate its prediction performance on \mathbf{D}_2 , based on which the weights $\mathbf{w} = \{w_1, \dots, w_K\}$ can be computed. Let $\beta_s^{(k)}$ be the sub-vector of $\beta^{(k)}$ representing the nonzero coefficients of model \mathcal{A}^k , and let $\mathbf{x}_s^{(k)} \in \mathbb{R}^{|\mathcal{A}^k|}$ be the corresponding subset of selected predictors. When p is large, the ARM weighting performs poorly for measuring the model deviation. One way to fix this is to add a non-uniform prior $e^{-\psi C_k}$ in the weighting computation

with

$$C_k = s_k \log \frac{ep}{s_k} + 2 \log(s_k + 2), \quad (4.7)$$

where s_k is the number of non-constant predictors for model k . The first term $s_k \log \frac{ep}{s_k}$ is an upper bound of $\log \binom{p}{s_k}$, which characterizes which model it is among the $\log \binom{p}{s_k}$ many possibilities. This is followed by

$$\binom{p}{s_k} = \frac{\prod_{j=0}^{s_k-1} (p-j)}{s_k!} \leq \frac{p^{s_k}}{s_k!} \leq \left(\frac{pe}{s_k}\right)^{s_k} \quad (4.8)$$

using Stirling's approximation. The second term in (4.7) represents the number of variables to be estimated. From an information-theoretic perspective, C_k can be regarded as an upper bound on the descriptive complexity of model \mathcal{A}^k . This concept plays a crucial role in model selection theory (Yang, 1999; Wang et al., 2014; Ye and Yang, 2019). Besides this interpretation, one can also treat $e^{-\psi C_k}$ as the prior probability assigned to the models from a Bayesian viewpoint. The constant $\psi > 0$ controls the relative importance of the prior weight on the final weights, which can be specified by the user. From a theoretical point of view, when ψ is bigger than 5.1, the complexity term is big enough to control the selection bias and results in minimax optimal estimations (Yang, 1999). But the bound 5.1 is more due to technical reasons. In practice, a smaller choice often works very well. Based on previous works (Nan and Yang, 2014; Ye et al., 2018; Ye and Yang, 2019) and our own numerical studies (Section 6 of the supplement), we found that $\psi = 1$ or 2 often delivers the best numerical results.

4.2 Weighting methods

The ARM weighting method for linear regression models is summarized in Algorithm 1.

Algorithm 1: ARM weighting procedure for linear regression.

- 1 Randomly split \mathbf{D} into a training set \mathbf{D}_1 and a test set \mathbf{D}_2 of equal size.
- 2 For each $\mathcal{A}^k \in \mathbb{S}$, fit a standard linear regression of y on $\mathbf{x}_s^{(k)}$ using the training set \mathbf{D}_1 and get the estimated regression coefficient $\hat{\boldsymbol{\beta}}_s^{(k)}$ and the estimated standard deviation $\hat{\boldsymbol{\sigma}}_s^{(k)}$.
- 3 For each \mathcal{A}^k , compute the prediction $\mathbf{x}_s^{(k)\top} \hat{\boldsymbol{\beta}}_s^{(k)}$ on the test set \mathbf{D}_2 .

- 4 Compute the weight w_k for each candidate model \mathcal{A}^k :

$$w_k = \frac{e^{-\psi C_k} (\hat{\boldsymbol{\sigma}}_s^{(k)})^{-n/2} \prod_{(\mathbf{x}_{s_i}^{(k)}, y_i) \in \mathbf{D}_2} \exp(-(\hat{\boldsymbol{\sigma}}_s^{(k)})^{-2} (y_i - \mathbf{x}_s^{(k)\top} \hat{\boldsymbol{\beta}}_s^{(k)})^2 / 2)}{\sum_{l=1}^K e^{-\psi C_l} (\hat{\boldsymbol{\sigma}}_s^{(l)})^{-n/2} \prod_{(\mathbf{x}_{s_i}^{(l)}, y_i) \in \mathbf{D}_2} \exp(-(\hat{\boldsymbol{\sigma}}_s^{(l)})^{-2} (y_i - \mathbf{x}_s^{(l)\top} \hat{\boldsymbol{\beta}}_s^{(l)})^2 / 2)}$$

for $k = 1, \dots, K$, where $C_k, k = 1, \dots, K$ is defined in (4.7).

- 5 Repeat the steps above (with random data splitting) L times to get $w_k^{(l)}$ for

$$l = 1, \dots, L, \text{ and get } w_k = \frac{1}{L} \sum_{l=1}^L w_k^{(l)}.$$

Weighting using ARM for logistic regression

The ARM weighting method for logistic regression models is similar. We summarize it in Algorithm 2.

Weighting using modified BIC for linear and logistic regression

Information criteria such as BIC can be used as alternative ways for computing the weights. Let ℓ_k be the maximized likelihood for model k . Recall that BIC

Algorithm 2: ARM weighting procedure for logistics regression.

- 1 Randomly split \mathbf{D} into a training set \mathbf{D}_1 and a test set \mathbf{D}_2 of equal size.
- 2 For each $\mathcal{A}^k \in \mathbb{S}$, fit a standard logistic regression of y on $\mathbf{x}_s^{(k)}$ using the data in \mathbf{D}_1

and get the estimated conditional probability function $\hat{p}^{(k)}(\mathbf{x}_s^{(k)})$, $k = 1, \dots, K$,

$$\hat{p}^{(k)}(\mathbf{x}_s^{(k)}) \equiv \Pr(Y = 1 | X_s^{(k)} = \mathbf{x}_s^{(k)}) = \exp(\mathbf{x}_s^{(k)\top} \hat{\boldsymbol{\beta}}_s^{(k)}) / (1 + \exp(\mathbf{x}_s^{(k)\top} \hat{\boldsymbol{\beta}}_s^{(k)})).$$

- 3 For each \mathcal{A}^k , evaluate $\hat{p}^{(k)}(\mathbf{x}_s^{(k)})$ on the test set \mathbf{D}_2 .
- 4 Compute the weight w_k for each model \mathcal{A}^k in the candidate models:

$$w_k = \frac{e^{-\psi C_k} \prod_{(\mathbf{x}_{s,i}^{(k)}, y_i) \in \mathbf{D}_2} \hat{p}^{(k)}(\mathbf{x}_{s,i}^{(k)})^{y_i} (1 - \hat{p}^{(k)}(\mathbf{x}_{s,i}^{(k)})^{1-y_i}}{\sum_{l=1}^K e^{-\psi C_l} \prod_{(\mathbf{x}_{s,i}^{(l)}, y_i) \in \mathbf{D}_2} \hat{p}^{(l)}(\mathbf{x}_{s,i}^{(l)})^{y_i} (1 - \hat{p}^{(l)}(\mathbf{x}_{s,i}^{(l)})^{1-y_i}}, \quad k = 1, \dots, K.$$

- 5 Repeat the steps above (with random data splitting) L times to get $w_k^{(l)}$ for

$$l = 1, \dots, L, \text{ and get } w_k = \frac{1}{L} \sum_{l=1}^L w_k^{(l)}.$$

is $I_k^{\text{BIC}} = -2 \log \ell_k + s_k \log n$. To accommodate the huge number of models, an extra term was added by Yang and Barron (1998) to reflect the additional price one needs to pay for searching through all the models. Including this extra term in the information criteria, we calculate the weights by a modified BIC (BIC-p) information criterion:

$$w_k = \exp(-I_k/2 - \psi C_k) / \sum_{l=1}^K \exp(-I_l/2 - \psi C_l), \quad k = 1, \dots, K. \quad (4.9)$$

5. SIMULATION

In order to study the performance of estimated F - and G -measures, we conduct simulations for several well-known variable selection methods (for both regression and classification) under various settings. We consider numerical ex-

5.1 Setting I: regression models

periments for both $n < p$ and $n \geq p$ cases, with specified structural feature correlation (i.e., independent/correlated). We also consider some special settings of the true coefficients such as decaying coefficients.

5.1 Setting I: regression models

For the regression case, the response Y is generated from the following model

$$Y = X\boldsymbol{\beta} + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$. To study how the estimation performance varies with the noise level σ^2 , we choose nine σ -values evenly spaced between 0.01 and 5. The predictors \mathbf{x}_i and the coefficient vector $\boldsymbol{\beta}$ are generated according to the following settings:

Example 1. $n = 200$, $p = 8$, $\boldsymbol{\beta} = (3, 1.5, 2, 0, 0, 0, 0, 0)^\top$. Predictors \mathbf{x}_i for $i = 1, \dots, n$ are generated as n i.i.d. observations from $N(0, \mathbf{I}_p)$.

Example 2. Same as Example 1 except $n = 1000$.

Example 3. $n = 200$, $p = 2000$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$, where $(\beta_1, \beta_2, \beta_3) = (3, 1.5, 2)$ and $(\beta_4, \dots, \beta_{2000})$ are zero. Predictors \mathbf{x}_i for $i = 1, \dots, n$ are sampled as n i.i.d. observations from $N(0, \mathbf{I}_p)$.

Example 4. $n = 200$, $p = 30$, components 1–5 of $\boldsymbol{\beta}$ are 10.5, components 6–10 are 5.5, components 11–15 are 0.5 and the rest are zero, so there are 15

5.1 Setting I: regression models

nonzero predictors, including five large ones, five moderate ones and five small ones. Predictors \mathbf{x}_i for $i = 1, \dots, n$ are generated from $X \sim N_p(0, \Sigma)$ with $\Sigma = (0.4^{|j-k|})_{p \times p}$, thus the pairwise correlation between X_j and X_k is $0.4^{|j-k|}$.

Example 5. $n = 200$, $p = 200$, components 1–5 of β are 10.5, components 6–10 are 5.5, components 11–15 are 0.5 and the rest are zero. Predictors \mathbf{x}_i for $i = 1, \dots, n$ are generated from $X \sim N_p(0, \Sigma)$. The covariance structure Σ is set as follows: the first 15 predictors (X_1, \dots, X_{15}) and the remaining 185 predictors (X_{16}, \dots, X_{200}) are independent. The pairwise correlation between X_j and X_k in (X_1, \dots, X_{15}) is $0.4^{|j-k|}$ with $j, k = 1, \dots, 15$. The pairwise correlation between X_j and X_k in (X_{16}, \dots, X_{200}) is $0.4^{|j-k|}$ with $j, k = 16, \dots, 200$.

We apply four penalized methods, Lasso, adaptive Lasso, MCP and SCAD to the data from Examples 1–5, and denoted by $\mathcal{A}^{\text{Lasso}}$, $\mathcal{A}^{\text{AdLasso}}$, \mathcal{A}^{MCP} and $\mathcal{A}^{\text{SCAD}}$ the resulting models, respectively. We use **glmnet** for computing $\mathcal{A}^{\text{Lasso}}$ and $\mathcal{A}^{\text{AdLasso}}$, and **ncvreg** for computing \mathcal{A}^{MCP} and $\mathcal{A}^{\text{SCAD}}$. Five-fold cross-validation is used for penalty parameter tuning in all these procedures. Because we know the true model $\mathcal{A}^* = \{j : \beta_j \neq 0\}$ in the simulation, we can report the true $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$ measures for each model-under-check $\mathcal{A}^0 \in \{\mathcal{A}^{\text{Lasso}}, \mathcal{A}^{\text{AdLasso}}, \mathcal{A}^{\text{MCP}}, \mathcal{A}^{\text{SCAD}}\}$. For comparison, we also compute estimated \hat{F} and \hat{G} using two different weighting methods, ARM and BIC-p (the modified BIC) with prior adjustment $\psi = 1$. The number of observations in the training set for computing the ARM weight is half of the sample size $\lfloor n/2 \rfloor$, and the

corresponding repetition time is 100.

All simulation cases are repeated for 100 times and the corresponding values are computed and averaged. We compare $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ with the true $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$ in Figure 1 for Example 1, and in Figures A1–A4 of the supplement for Examples 2–5. Overall, $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ using ARM and BIC-p weighting can well reflect the trends of $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$ in the sense that, both the true curves and the estimated curves trend down as σ^2 increases. And the estimation accuracy drops as σ^2 increases. The estimated $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ properly reflect the true performance of a given \mathcal{A}^0 . For example, in Figures A2, A3 and A4, we see that the performance of Lasso deteriorates significantly as σ^2 increases, due to the fact that it tends to over-select variables under higher noise levels. In contrast, adaptive Lasso, MCP and SCAD have more robust performance against the high noise. $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ can correctly reflect these aforementioned facts. From the results, we find that MCP is the best performer with the highest true/estimated F - and G -measures in Examples 2–5, while adaptive Lasso is the best performer in Example 1.

By comparing Figures 1 and A1, we see that the sample size influences the estimation performance: large samples produce more accurate $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$. Gains in the estimation accuracy from increased sample sizes are due to the fact that more information results in better assigned weights on the candidate models.

5.1 Setting I: regression models

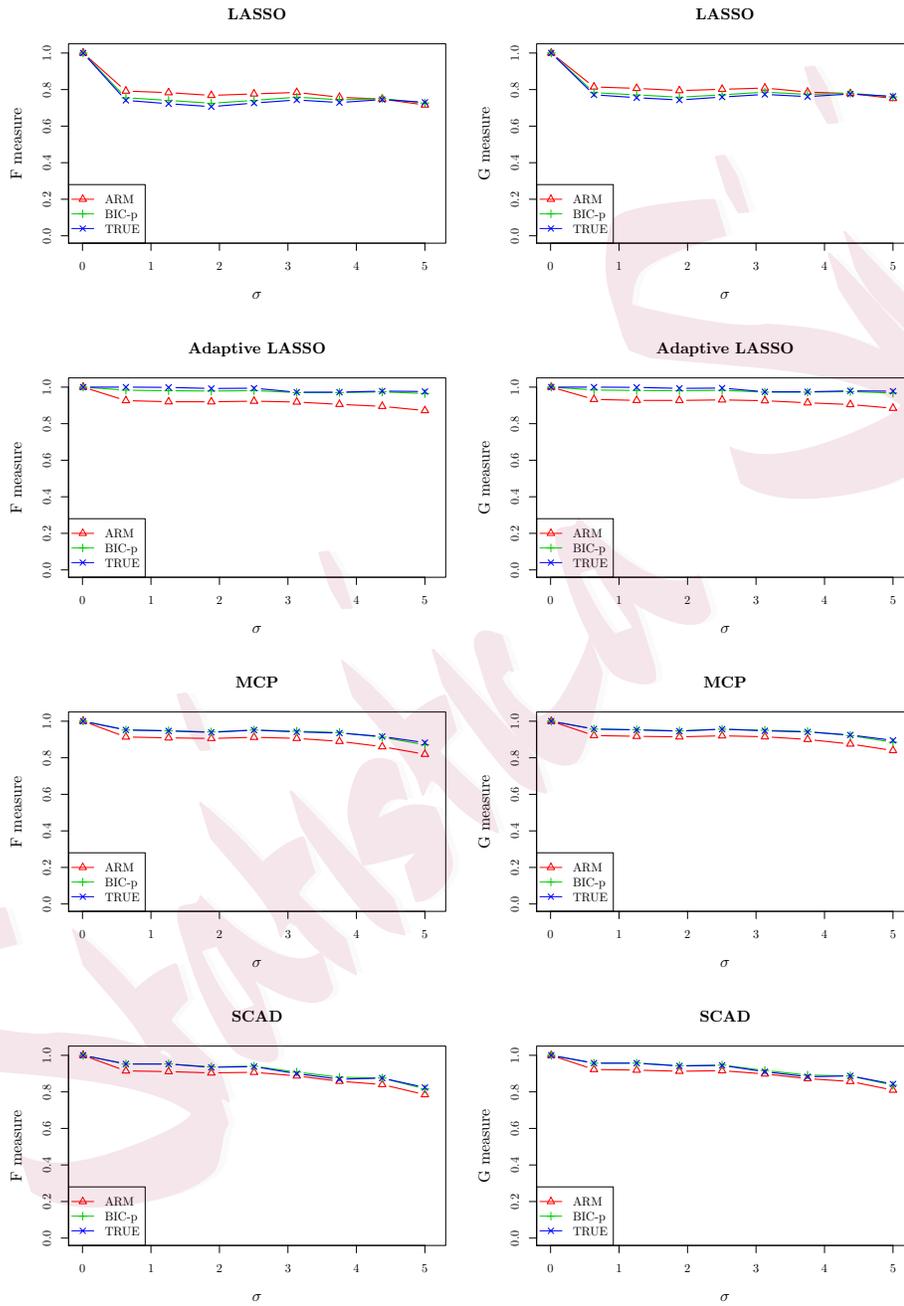


Figure 1: Regression case (Example 1).

5.2 Setting II: classification models

In Figure A4, the over-estimation in Adaptive Lasso, SCAD and MCP, when σ is large, is due to the fact that highly weighted candidate models miss several small coefficients variables, which is caused by the decaying coefficients and worsened by correlation between the variables. While for Lasso, when σ is small, PAVI can find good candidate models to put high weights on, thus the estimation is good; when σ is larger, the candidate models with high weights miss several true variables. At the same time, Lasso chooses more redundant variables when σ becomes larger. Therefore, the precision is under-estimated, so does the F -measure.

5.2 Setting II: classification models

For the classification case, we randomly generate n i.i.d observations $\{y_i, \mathbf{x}_i\}_{i=1}^n$. Each binary response $y_i \in \{0, 1\}$ is generated according to the Bernoulli distribution with the conditional probability $\Pr(Y = 1|X = \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}$. The explanatory variables X and the coefficient vector $\boldsymbol{\beta}$ are set under the same configurations as in Examples 1–5.

The absolute differences between the true and estimated measures,

$$d_F = |\widehat{F}(\mathcal{A}^0) - F(\mathcal{A}^0)| \quad \text{and} \quad d_G = |\widehat{G}(\mathcal{A}^0) - G(\mathcal{A}^0)|$$

are used to evaluate the estimation performance, where the smaller d_F and d_G , the better the estimation performance.

All simulation cases are repeated for 100 times and the corresponding $F(\mathcal{A}^0)$,

5.2 Setting II: classification models

$G(\mathcal{A}^0)$, $\widehat{F}(\mathcal{A}^0)$, $\widehat{G}(\mathcal{A}^0)$, d_F and d_G values are computed and averaged. The results are summarized in Table 1 for Example 1, and in Tables A1–A4 of the supplement for Examples 2–5. The standard errors are also shown (in the parentheses). As we can see from those tables, d_F and d_G are generally small, which indicates that the estimated $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ are good approximations to the true $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$, respectively. The estimated $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ can reflect the true advantage of a given variable selection method. For example, in Table 1, and Tables A1–A4, we can see that adaptive Lasso, MCP and SCAD have better variable selection performance than Lasso according to their larger true values of $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$. The estimated $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ can correctly reflect these differences in performance.

Our estimation method can still perform very well under the high-dimensional setting, which can be seen from the small d_F and d_G in Table A2. However, the results from Tables 4 and 5 show that the decaying coefficients and feature correlation make the estimation of $\widehat{F}(\mathcal{A}^0)$ and $\widehat{G}(\mathcal{A}^0)$ more difficult. In these two cases, BIC-p methods tend to over-estimate $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$ for MCP and SCAD, while ARM tends to under-estimate $F(\mathcal{A}^0)$ and $G(\mathcal{A}^0)$ for Lasso and adaptive Lasso.

The over-estimation problem of the BIC-p method mainly comes from over-estimation of the recall part. The final model selected by SCAD misses several true variables, thus the true recall is very small. However, if one uses the heavily

5.2 Setting II: classification models

weighted candidate models that miss several true variables in the PAVI calculation, the recall would be over-estimated.

For SCAD and ARM combination, using the heavily weighted models that miss several true variables in PAVI will give us over-estimation of the recall and under-estimation of the precision, while these two effects cancel each other to some degree.

The under-estimation by ARM methods mainly comes from the under-estimation of the precision part, while the estimated recall is close to (slightly over-estimating) the true recall. Lasso tends to miss true variables and over-select redundant variables in the examples. Thus, the true precision of Lasso is small.

For Lasso and BIC combination, using the heavily weighted models that miss several true variables with small coefficients in PAVI computing will give us over-estimation of the recall and under-estimation of the precision, while these two effects cancel each other to some degree.

Both issues are mainly caused by the fact that the candidate models with large weights could not recover all the variables with small true coefficients, and the problem is further worsened by the existence of high correlation among the features.

Table 1: Classification case (Example 1).

	F	G	d_F	d_G
Lasso				
True	0.670 (0.010)	0.712 (0.009)		
ARM	0.711 (0.009)	0.747 (0.007)	0.046 (0.003)	0.039 (0.002)
BIC-p	0.687 (0.010)	0.726 (0.008)	0.017 (0.002)	0.014 (0.001)
AdLasso				
True	0.944 (0.009)	0.949 (0.008)		
ARM	0.899 (0.004)	0.908 (0.004)	0.066 (0.003)	0.060 (0.003)
BIC-p	0.946 (0.007)	0.950 (0.007)	0.018 (0.002)	0.016 (0.001)
MCP				
True	0.968 (0.009)	0.971 (0.008)		
ARM	0.903 (0.005)	0.913 (0.004)	0.079 (0.003)	0.072 (0.002)
BIC-p	0.961 (0.007)	0.965 (0.006)	0.019 (0.002)	0.017 (0.001)
SCAD				
True	0.902 (0.012)	0.911 (0.010)		
ARM	0.881 (0.006)	0.892 (0.006)	0.054 (0.003)	0.050 (0.003)
BIC-p	0.911 (0.010)	0.919 (0.009)	0.018 (0.002)	0.016 (0.001)

6. REAL DATA

In this section, we apply PAVI using candidate models from several model selection methods to gene expression data for cancer-related biomarker identification. The biomarker selection process is usually under high-dimensional, small-sample, and high-noise setting involving highly-correlated genes (Golub et al., 1999; West et al., 2001). As such, the sets of genes identified may be subject to substantial changes due to small perturbations in the data (Baggerly et al., 2004; Henry and Hayes, 2012). Here we use \hat{F} and \hat{G} to evaluate such selection uncertainty.

Our goal is to provide a serious and careful analysis of the outcomes of

6.1 Data description

Table 2: Summary of Colon, Leukemia, Prostate.

Data	n	n_1 ($y = 1$)	n_2 ($y = 0$)	p (number of genes)	Data source
Colon	62	40	22	2000	Alon et al. (1999)
Leukemia	72	25	47	7129	Golub et al. (1999)
Prostate	102	52	50	12600	Singh et al. (2002)

several variable selection methods from multiple angles to understand the key issues of interest. One may wonder if any strong statement can be said because no one knows the truth. We hope our analysis provides strong enough evidence that the estimated F and G values yield valuable information.

6.1 Data description

We consider three well-studied benchmark cancer datasets: `Colon` (Alon et al., 1999), `Leukemia` (Golub et al., 1999) and `Prostate` (Singh et al., 2002). Table 2 provides a brief summary.

6.2 Methods/models to be examined

On these three datasets, we compare the variable selection performance of four commonly used penalization methods: Lasso, adaptive Lasso, MCP and SCAD. We first obtain the final model \mathcal{A}^0 for each method (the tuning parameter λ is selected using five-fold cross-validation). Then we use PAVI to estimate $\hat{F}(\mathcal{A}^0)$ and $\hat{G}(\mathcal{A}^0)$ with two weighting schemes, ARM and BIC-p. The whole procedure is repeated 100 times to average out randomness in the tuning parameter

6.2 Methods/models to be examined

selection, and the averages of $\widehat{F}(\mathcal{A}^0)$, $\text{sd}(\widehat{F}(\mathcal{A}^0))$ and $\widehat{G}(\mathcal{A}^0)$, $\text{sd}(\widehat{G}(\mathcal{A}^0))$ are summarized in Tables 3, 4 and A5. For comparison, we also include several other models studied in the existing literature. Specifically, we consider Leung and Hung, 2010 (L10), Yang and Song, 2010 (Y10), Chandra and Gupta, 2011 (C11) and Lee and Leu, 2011 (L11) for Colon, Leung and Hung, 2010 (L10), Yang and Song, 2010 (Y10), and Ji et al., 2011 (J11; two kinds of models are provided via different importance criterion in this work, denoted by J11¹ and J11² hereafter respectively) for Leukemia, and Leung and Hung, 2010 (L10) and Sharma et al., 2012 (S12) for Prostate.

Y10, J11 and S12 used linear-based variable selection techniques without initial variable screening. Specifically, Y10 used the probit regression model; J11 used the linear kernel support vector classifier (SVC); S12 used the linear discriminant analysis (LDA) technique with nearest centroid classifier (NCC). In contrast, L10, C11 and L11 used nonparametric variable selection techniques: L10 used the support vector machine (SVM); C11 used the naïve Bayes classifier (NBC) and SVM; L11 used SVM. In addition, we consider the Importance Screening method (ImpS) by Ye et al. (2018), which uses a sparsity oriented importance learning for variable screening.

Table 3: Estimated F - and G -measures and standard deviations for Colon. L10 has numerically zero \hat{F} and \hat{G} values (bolded in the Table).

	ARM				BIC-p			
	F	$sd.F$	G	$sd.G$	F	$sd.F$	G	$sd.G$
Lasso	0.147	0.024	0.280	0.022	0.205	0.066	0.332	0.058
AdLasso	0.194	0.165	0.255	0.211	0.309	0.191	0.361	0.209
MCP	0.349	0.045	0.459	0.035	0.460	0.130	0.544	0.093
SCAD	0.149	0.032	0.274	0.039	0.211	0.074	0.331	0.071
ImpS	0.524	0.081	0.596	0.065	0.656	0.176	0.698	0.118
L11	0.111	0.110	0.175	0.175	0.112	0.105	0.157	0.151
Y10	0.103	0.017	0.233	0.018	0.146	0.048	0.276	0.047
C11	0.184	0.020	0.317	0.022	0.223	0.076	0.333	0.082
L10	0.000							

6.3 Results

The estimated \hat{F} and \hat{G} of each model on Colon, Leukemia and Prostate are reported in Tables 3, 4 and A5 (in the supplement), respectively. We find that ImpS achieves almost the largest estimated \hat{F} and \hat{G} on all three data sets. L10 has basically zero \hat{F} and \hat{G} for Colon and Prostate. J11¹ and J11² has basically zero \hat{F} and \hat{G} for Leukemia. (These cases are bolded in Tables 3, 4 and A5.) This suggests that, from a logistic regression modeling perspective, they may have chosen “wrong” variables and they have very low recalls or precisions.

6.4 Are the zero \hat{F} and \hat{G} values too harsh for the methods?

It is striking that the \hat{F} and \hat{G} values for some selections are numerically zero, which seems rather extreme. Does this mean those models are truly poor or rather our performance assessment methodology fails? We would like to examine

6.4 Are the zero \widehat{F} and \widehat{G} values too harsh for the methods?

Table 4: Estimated F - and G -measures and standard deviations for Leukemia.

J11¹ and J11² have numerically zero \widehat{F} and \widehat{G} values (bolded in the Table).

	ARM				BIC-p			
	F	$sd.F$	G	$sd.G$	F	$sd.F$	G	$sd.G$
Lasso	0.083	0.025	0.206	0.026	0.079	0.012	0.203	0.014
AdLasso	0.323	0.044	0.432	0.031	0.322	0.039	0.434	0.033
MCP	0.168	0.170	0.221	0.210	0.061	0.089	0.078	0.108
SCAD	0.094	0.028	0.220	0.028	0.090	0.013	0.216	0.015
ImpS	0.525	0.065	0.591	0.042	0.573	0.129	0.636	0.102
J11 ¹	0.000							
J11 ²	0.000							
Y10	0.108	0.014	0.236	0.009	0.105	0.002	0.233	0.012
L10	0.212	0.180	0.265	0.224	0.336	0.089	0.419	0.110

the matter from three perspectives.

6.4.1 First perspective: the labels of the selected genes

First, let us examine the labels of the selected genes. We obtain the selected genes in the literature. And we use five-fold cross-validation in penalty parameter tuning to obtain selected genes for the penalized regression models. In Tables A6, A7 and A8, the results show that the genes selected by L10 (Colon and Prostate), J11¹ and J11² (Leukemia) are mostly not supported by other models. More specifically, the choices of variables by L10, J11¹ and J11² in those cases share zero, one or at most two genes with the other methods, respectively.

(These cases are underlined in Tables A6, A7 and A8.)

6.4 Are the zero \hat{F} and \hat{G} values too harsh for the methods?

6.4.2 Second perspective: predictive accuracy

Secondly, we would like to examine the issue from a predictive accuracy perspective. We randomly split the dataset into 4/5 observations for training and 1/5 observations for testing. We fit the SVM models with those selected genes on the training data using **kernlab** (Karatzoglou et al., 2004) and evaluate the predictive accuracy on the testing data. The whole procedure is repeated 100 times and the averaged classification accuracy and “standard errors” (w.r.t. the permutations) are recorded in Table 5. Alternatively, we may consider the parametric models. We fit the logistic regression with the genes selected (in Table 5). We find that L10, J11¹ and J11² have worse predictive accuracy (bolded in Table 5) compared with the simpler model by ImpS, which adds evidence to support the validity of their low \hat{F} and \hat{G} values.

6.4.3 Third perspective: traditional model fitting

For the third perspective, we investigate the AIC, BIC, and deviance measures. When comparing models fitted by maximum likelihood to the same data, the smaller the AIC or BIC value, the better the model, from their respective stand points.

From Table 6, the model for **Colon** with zero \hat{F} and \hat{G} values also has relatively large AIC, BIC and deviance values (bolded in the Table) compared to the models with large \hat{F} and \hat{G} values. The results are similar for the other

6.4 Are the zero \widehat{F} and \widehat{G} values too harsh for the methods?

Table 5: Comparisons of classification accuracy on Colon, Leukemia, and Prostate using logistic regression and SVM, respectively.

Logistic Model					
Colon		Leukemia		Prostate	
ImpS	86.3 (0.8)	ImpS	97.1 (0.3)	ImpS	94.0 (0.4)
Lasso	80.0 (1.0)	Lasso	99.8 (0.1)	Lasso	97.0 (0.4)
AdLasso	85.5 (0.8)	AdLasso	93.9 (0.5)	AdLasso	99.8 (0.1)
MCP	85.1 (0.8)	MCP	99.5 (0.1)	MCP	98.7 (0.2)
SCAD	84.3 (0.8)	SCAD	97.9 (0.3)	SCAD	97.1 (0.2)
L11	80.4 (0.8)	J11 ¹	89.4 (0.8)	S12	96.5 (0.4)
Y10	90.9 (0.9)	J11 ²	89.8 (0.7)	L10	59.0 (0.8)
C11	79.6 (1.0)	Y10	91.2 (0.7)		
L10	83.0 (0.9)	L10	95.5 (0.4)		
SVM Model					
Colon		Leukemia		Prostate	
ImpS	84.0 (0.9)	ImpS	97.6 (0.3)	ImpS	95.3 (0.4)
Lasso	75.8 (0.9)	Lasso	99.1 (0.2)	Lasso	96.3 (0.4)
AdLasso	79.0 (0.9)	AdLasso	95.8 (0.4)	AdLasso	96.6 (0.3)
MCP	83.1 (1.1)	MCP	99.0 (0.2)	MCP	97.1 (0.3)
SCAD	86.0 (0.9)	SCAD	99.1 (0.2)	SCAD	96.4 (0.3)
L11	79.0 (1.1)	J11 ¹	88.6 (0.8)	S12	95.5 (0.4)
Y10	78.3 (1.0)	J11 ²	87.4 (0.9)	L10	59.3 (0.9)
C11	77.1 (0.9)	Y10	90.2 (0.6)		
L10	72.4 (0.9)	L10	92.2 (0.6)		

Table 6: Estimated AIC, BIC and deviance for Colon, Leukemia and Prostate.

	Colon			Leukemia			Prostate				
	AIC	BIC	Dev.	AIC	BIC	Dev.	AIC	BIC	Dev.		
Lasso	26.0	53.6	0.0	56.0	119.7	0.0	62.0	143.3	0.0		
AdLasso	34.9	49.8	20.9	12.0	25.6	0.0	22.0	50.8	0.0		
MCP	32.1	44.9	20.1	16.0	34.2	0.0	16.0	36.9	0.0		
SCAD	26.0	53.6	0.0	48.0	102.6	0.0	38.0	87.8	0.0		
ImpS	35.5	44.1	27.5	8.0	17.1	0.0	12.0	27.7	9.4		
L11	51.4	70.5	33.4	J11 ¹	20.0	42.7	0.0	S12	36.1	49.2	26.1
Y10	40.0	82.5	0.0	J11 ²	18.0	38.4	0.0	L10	140.1	158.5	126.1
C11	45.2	68.6	23.2	Y10	38.0	81.2	0.0				
L10	48.6	63.5	34.6	L10	10.0	21.3	0.0				

two data sets, except that the deviance values for `Leukemia` are extremely small due to the easy classification nature of the data.

In summary, we see that the low (near zero) \hat{F} and \hat{G} values for the above investigated sets of selected genes are supported from the three perspectives. Our PAVI approach provides a valid tool for checking the reliability and reproducibility of a given set of selected variables when the true model is not known. To be fair, we want to emphasize that the poor \hat{F} and \hat{G} values of some of the selection methods are based on the logistic regression perspective, although Table 5 seems to suggest that logistic regression works at least as well as SVM.

7. CONCLUSION

There are many variable selection methods, but so far most of investigations on their behaviors are limited to theoretical studies and somewhat scattered simulation results, which may have little to do with a specific dataset at hand. There is a severe lack of valid performance measures that are computable based on data alone. This leads to the pessimistic view that “For real data, nothing can be said strongly about which method is better for describing the data generation mechanism since no one knows the truth.” Sound implementable variable selection diagnostic tools can shed a positive light on the matter.

Nan and Yang (2014) proposed an approach to gain insight on how many variables are likely missed and how many are not quite justifiable for an outcome

of a variable selection process. In real applications, it is often of interest and important to summarize the two types of selection errors into a single measure to characterize the behavior of a variable selection method. Due to this reason, F - and G -measures are gaining popularity in model selection literature. If we are given a data set for which several model selection methods are considered, prior to this work, the available model diagnostic tools can only tell us (a) which methods are more unstable; (b) how many terms are likely missed or unsupported. This information, unlike the F - and G -measures, may not be enough to give one a good sense of the overall model selection performance. In this paper, we have advanced the line of research on model selection diagnostics by providing a valid estimation of F - and G -measures.

We have proved that the estimated F - and G -measures are uniformly consistent as long as the weighting is weakly consistent. The simulation results clearly show that the \hat{F} and \hat{G} values based on our PAVI approach nicely characterize the overall performance of the model selection outcomes. The information can be utilized for comparing different methods for the data at hand.

We have used three real data examples to demonstrate the utility of our PAVI methodology. There have been many variable selection results reported in the literature on these data sets. A careful study with multiple perspectives has provided strong evidence to suggest that some of the variable selection outcomes may be far away from the best set of variables to use for logistic regression or

SVM with the given information.

Supplementary Materials

Text document: Supplementary material for “Performance Assessment of High-dimensional Variable Identification”. (.pdf file)

Acknowledgment

The authors thank the Editor, an associate editor, and two anonymous referees for their helpful comments and suggestions. The work of Yi Yang is partially supported by NSERC RGPIN-2016-05174 and FRQNT NC-205972.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pp. 267–281. Budapest: Akadémiai Kiadó.
- Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences* 96(12), 6745–6750.
- Baggerly, K. A., J. S. Morris, and K. R. Coombes (2004). Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* 20(5), 777–785.

REFERENCES

- Billsus, D. and M. J. Pazzani (1998). Learning collaborative information filters. In *International Conference on Machine Learning*, Volume 98, pp. 46–54. Morgan Kaufmann Publishers.
- Breheny, P. and J. Huang (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics* 5(1), 232.
- Breiman, L. (1996a). Bagging predictors. *Machine Learning* 24(2), 123–140.
- Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* 24(6), 2350–2383.
- Buckland, S., K. Burnham, and N. Augustin (1997). Model selection: an integral part of inference. *Biometrics* 53(2), 603–618.
- Chandra, B. and M. Gupta (2011). An efficient statistical feature selection approach for classification of gene expression data. *Journal of Biomedical Informatics* 44(4), 529–535.
- Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 158(3), 419–466.
- Chen, L., P. Giannakouros, and Y. Yang (2007). Model combining in factorial data analysis. *Journal of Statistical Planning and Inference* 137(9), 2920–2934.
- Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 57(1), 45–97.
- Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456), 1348–1360.
- Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space.

REFERENCES

Statistica Sinica 20(1), 101.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1), 1.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, and M. A. Caligiuri (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286(5439), 531–537.

Henry, N. L. and D. F. Hayes (2012). Cancer biomarkers. *Molecular Oncology* 6(2), 140–146.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science* 14(4), 382–401.

Ji, G., Z. Yang, and W. You (2011, Nov). PLS-based gene selection and identification of tumor-specific genes. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41(6), 830–841.

Karatzoglou, A., A. Smola, K. Hornik, and A. Zeileis (2004). kernlab - An S4 package for kernel methods in R. *Journal of Statistical Software* 11(1), 1–20.

Lai, R. C. S., J. Hannig, and T. C. M. Lee (2015). Generalized fiducial inference for ultrahigh-dimensional regression. *Journal of the American Statistical Association* 110(510), 760–772.

Lee, C.-P. and Y. Leu (2011). A novel hybrid feature selection method for microarray data analysis. *Applied Soft Computing* 11(1), 208–213.

Leung, G. and A. R. Barron (2006, Aug). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* 52(8), 3396–3410.

REFERENCES

- Leung, Y. and Y. Hung (2010). A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 7(1), 108–117.
- Lim, C. (2011). *Modeling High Dimensional Data: Prediction, Sparsity, and Robustness*. Ph. D. thesis, University of California, Berkeley.
- Lim, C. and B. Yu (2016). Estimation stability with cross-validation (ESCV). *Journal of Computational and Graphical Statistics* 25(2), 464–492.
- McNutt, M. (2014). Raising the bar. *Science* 345(6192), 9.
- Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34(3), 1436–1462.
- Nan, Y. and Y. Yang (2014). Variable selection diagnostics measures for high-dimensional regression. *Journal of Computational and Graphical Statistics* 23(3), 636–656.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6(2), 461–464.
- Sharma, A., S. Imoto, and S. Miyano (2012). A top-r feature selection algorithm for microarray gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9(3), 754–764.
- Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D’Amico, and J. P. Richie (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1(2), 203–209.
- Stodden, V. (2015). Reproducing statistical results. *Annual Review of Statistics and Its Application* 2,

REFERENCES

1–19.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 58(1), 267–288.

Wang, Z., S. Paterlini, F. Gao, and Y. Yang (2014). Adaptive minimax regression estimation over sparse lq-hulls. *The Journal of Machine Learning Research* 15(1), 1675–1711.

West, M., C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences* 98(20), 11462–11467.

Yang, A.-J. and X.-Y. Song (2010). Bayesian variable selection for disease classification using gene expression data. *Bioinformatics* 26(2), 215–222.

Yang, Y. (1999). Model selection for nonparametric regression. *Statistica Sinica* 9(2), 475–499.

Yang, Y. (2000). Adaptive estimation in pattern recognition by combining different procedures. *Statistica Sinica* 10(4), 1069–1090.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association* 96(454), 574–588.

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics* 35(6), 2450–2473.

Yang, Y. and A. R. Barron (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory* 44(1), 95–116.

Ye, C. and Y. Yang (2019). High-dimensional adaptive minimax sparse estimation with interactions.

REFERENCES

IEEE Transactions on Information Theory 65(9), 5367–5379.

Ye, C., Y. Yang, and Y. Yang (2018). Sparsity oriented importance learning for high-dimensional linear regression. *Journal of the American Statistical Association* 113(524), 1797–1812.

Yuan, Z. and Y. Yang (2005). Combining linear regression models: when and how? *Journal of the American Statistical Association* 100(472), 1202–1214.

Zhang, C.-H. (2010, 04). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38(2), 894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101(476), 1418–1429.

Yanjia Yu, School of Statistics, University of Minnesota

E-mail: yuxxx748@umn.edu

Yi Yang, Department of Mathematics and Statistics, McGill University

E-mail: yi.yang6@mcgill.ca

Yuhong Yang, School of Statistics, University of Minnesota

E-mail: yangx374@umn.edu