

Using SVM Weight-Based Methods to Identify Causally Relevant and Non-Causally Relevant Variables

Alexander Statnikov¹, Douglas Hardin^{1,2}, Constantin Aliferis^{1,3}

¹Department of Biomedical Informatics, ²Department of Mathematics, ³Department of Cancer Biology, Vanderbilt University, Nashville, TN 37232, USA.

Abstract: We conducted simulation experiments to study SVM weight-based ranking and variable selection methods using two network structures that are often encountered in biological systems and are likely to occur in many other settings as well. We attempted to recover both causally and non-causally relevant variables using SVM weight-based methods under a variety of experimental settings (data-generating network, noise level, sample size, and SVM penalty parameter). Our experiments show that SVMs produce excellent classifiers that often assign higher weights to irrelevant variables than to the relevant ones. Likewise, the application of the recursive variable selection technique SVM-RFE, does not remedy this problem. More importantly, we found that when it comes to identifying causally relevant variables, SVM weight-based methods can fail by assigning higher weights or selecting (in the context of SVM-RFE) variables that are relevant but non-causally so. Furthermore, even irrelevant variables can have higher weights or can be selected more frequently than the causally relevant ones. We show that this problem is not linked to the specific variable selection techniques studied but rather that the maximum margin inductive bias, as typically employed by SVM-based methods, is locally causally inconsistent. New SVM methods may be needed to address this issue and this is an exciting and challenging area of research.

Introduction and Background

Variable selection is one of the most important areas of machine learning [6,9], especially when it comes to analysis, modeling, and discovery from high-dimensional datasets. In addition to the promise of cost-effectiveness, two major goals of variable selection are to improve the prediction performance of the predictors and to provide a better understanding of the underlying process that generated the data [6]. To this end, variable selection is often used to derive insights in the *causal structure* of the data-generating process. For example, in biology and medicine, biomarker discovery is performed not simply to derive predictive and diagnostic models but also to better understand the factors that cause disease, determine its progression, and to identify the members of the relevant molecular pathways.

In the context of variable selection, it is common to categorize variables as *relevant* (ones that are conditionally dependent on the response variable given any subset of variables) and *irrelevant* (ones that are independent of the response variable) [9]. In causal networks, relevant variables have an undirected path to the response variable, while irrelevant variables do not [13]. In all experiments in the present paper, we further divide the class of relevant variables into *causally relevant* (ones that are causing the response variable) and *non-causally relevant* (all other relevant variables).

A recent breakthrough in variable selection is the development of Support Vector Machine (SVM) weight-based methods that scale up to datasets with many thousands of variables and as few as dozens of samples [7,11]. These methods achieve the first goal of variable selection, i.e. they often yield variables that are more predictive than the ones output by other variable selection techniques or the full (unreduced) variable set [6,7]. However, the extent to which the second goal of variable selection is achieved by the SVM weight-based methods (i.e., whether we get insights in the causal structure) has not received much attention in the literature yet. An exception is the work in [8] that provided a theoretical characterization of the linear SVM-based variable selection and suggested that (i) the irrelevant variables will be given a zero weight by a linear SVM in the sample limit, and (ii) the linear SVM may assign zero weight to causally relevant variables¹ and nonzero weight to non-causally relevant variables.

In the present paper, we conducted simulation experiments to study SVM weight-based ranking and variable selection methods using two network structures that are often encountered in biological systems and are likely to occur in many other settings as well. We attempted to recover both causally and non-causally relevant variables using SVM weight-based methods under a variety of experimental settings (data-generating network type, different number of relevant and irrelevant variables in the data-generating network, noise level, sample size, and SVM

¹ “Causally relevant” variables were defined in that study as members of the local causal neighborhood of the response variable.

penalty parameter). A comprehensive visualization of the results of all experiments is provided in the online supplement, available from <http://www.dsl-lab.org/supplements/NIPS2006/>.

In brief, our experiments show that SVMs can produce excellent classifiers that often assign higher weights² to irrelevant variables than to the relevant ones. Likewise, the application of the recursive variable selection technique SVM-RFE [7], does not remedy this problem. More importantly, we found that when it comes to identifying causally relevant variables, SVM weight-based methods can fail by assigning higher weights or selecting (in the context of SVM-RFE) variables that are relevant but non-causally so. Furthermore, even irrelevant variables can have higher weights or can be selected more frequently than the causally relevant ones. These results are corroborated by a theoretical analysis as well as recent research employing high-fidelity re-simulations in biological and medical domains [1-3]. Thus, available empirical evidence so far suggests that *causal interpretation of current state-of-the-art SVM variable selection results must be conducted with great caution by practitioners*. We show that this problem is not linked to the specific variable selection techniques studied but rather that the maximum margin inductive bias, as typically employed by SVM-based methods, is *locally causally inconsistent*. This means that in some distributions the same SVM weights can be assigned to variables inside and outside the local causal neighborhood of the response variable. This may occur even when neither the Causal Markov Condition nor Faithfulness are violated (which implies that causal discovery is feasible via standard causal discovery algorithms) [12]. Since any locally causally inconsistent procedure will also be globally inconsistent, broadly speaking, SVM weights cannot be used for learning causality in a sound way. New SVM methods may be needed to address this issue and this is an exciting and challenging area of research.

Simulation Experiments

1. Data simulation. We considered two types of network structures shown in Figures 1 and 2. In the *first* type of network structure (Figure 1),

- Y is a binary variable with $P(Y=0) = 1/2$ and $P(Y=1) = 1/2$. Y is used only for data generation and is hidden from the learner afterwards (i.e., Y is not present in the dataset analyzed by the algorithms).
- $\{X_i\}_{i=1,\dots,N}$ are binary variables with $P(X_i=0|Y=0) = q$ and $P(X_i=1|Y=1) = q$, where q is a fixed constant as described below.
- $\{Z_i\}_{i=1,\dots,M}$ are independent binary variables with $P(Z_i=0) = 1/2$ and $P(Z_i=1) = 1/2$.
- T is a binary response variable with $P(T=0|X_1=0) = 0.95$ and $P(T=1|X_1=1) = 0.95$.

In this network structure, variable X_1 is the only causally relevant one. Variables $\{X_i\}_{i=2,\dots,N}$ are also relevant but non-causally so. Variables $\{Z_i\}_{i=1,\dots,M}$ are irrelevant. In the simulation experiments, we used $q=0.95$ (we call the resulting network “**1a**”) and $q=0.99$ (“**1b**”). Figure 3 provides an illustration of this network structure in the biological pathways produced by Ariadne Genomics PathwayStudio software version 4.0 (<http://www.ariadnegenomics.com/>). *kras* is one of many proteins that is implicated for the adrenal gland carcinoma and corresponds to variable X_1 in this network structure. *SOS1* (corresponds to Y) is directly upstream of *kras* and it is regulating many proteins that may be strongly correlated with each other. Many similar examples exist in biological and other settings.

In the *second* type of network structure (Figure 2),

- $\{X_i\}_{i=1,\dots,N}$ are independent binary variables with $P(X_i=0) = 1/2$ and $P(X_i=1) = 1/2$.
- $\{Z_i\}_{i=1,\dots,M}$ are independent binary variables with $P(Z_i=0) = 1/2$ and $P(Z_i=1) = 1/2$.
- Y is a “synthesis variable” with the following function: $Y = 1/N \sum_{i=1}^N X_i$.
- T is a binary response variable defined as $T = \text{sign}\left(\sum_{i=1}^N v_i X_i - N/4\right)$, where v_i 's are generated from the uniform random $U(0,1)$ distribution and are fixed for all experiments.

In this network structure, variables $\{X_i\}_{i=1,\dots,N}$ are causally relevant. Variable Y is relevant but non-causally so. Variables $\{Z_i\}_{i=1,\dots,M}$ are irrelevant. We will call this network “**2**”. Figure 4 shows several biological instances of this example structure: a pathway where three regulatory genes are mutually involved in the coordinate regulation of 28-138 genes [10]. This example is also common in many biological and other settings.

² By “weight” here and elsewhere in the paper we mean “absolute value of the SVM weight”.

From the above three networks (1a, 1b, 2), we generated 30 training random samples of sizes = {100, 200, 500, 1000} for different values of N (number of all relevant variables³) = {10, 100} and M (number of irrelevant variables) = {10, 100, 1000}. We also generated testing random samples of size 5000 for all three above networks and different values of N and M. Once the datasets were generated, we added noise (both to the training and testing datasets) to simulate random measurement errors. The noise model was implemented as follows: replace X% of each variable values with values randomly sampled from the distribution of that variable in simulated data. We experimented with {0%, 1%, 10%} noise.

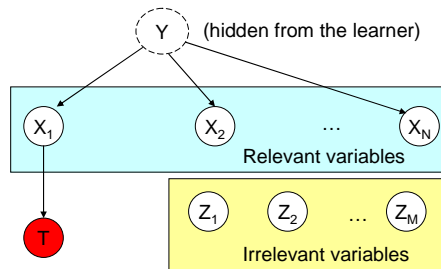


Figure 1. The *first* type of network structure. X_1 is the only causally relevant variable.

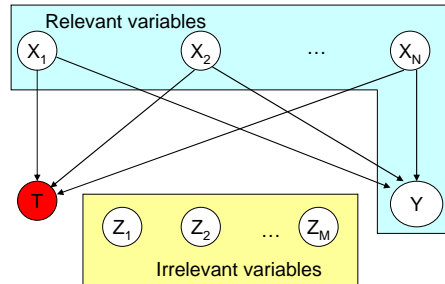


Figure 2. The *second* type of network structure. X_1, \dots, X_N are causally relevant variables.

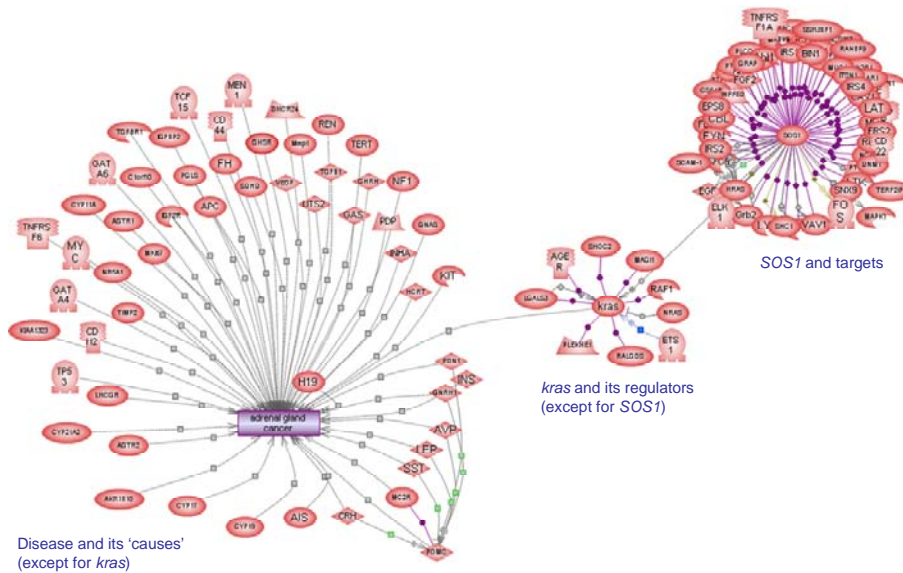
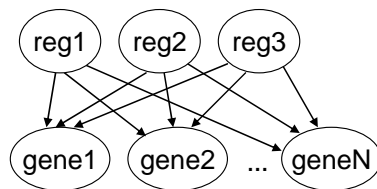


Figure 3. Real-world example of the *first* type of network structure. Adrenal gland cancer pathway produced by Ariadne Genomics PathwayStudio software version 4.0 (<http://www.ariadnegenomics.com/>).



Regulators			Number of genes (N)
reg1	reg2	reg3	
HNF1A	FOXA2	HNF4A	62
HNF1A	HNF4A	HNF6	55
HNF1A	HNF4A	CREB1	138
HNF1A	HNF4A	USF1	50
FOXA2	HNF4A	HNF6	92
FOXA2	HNF4A	CREB1	28
FOXA2	HNF4A	USF1	34
HNF4A	HNF6	CREB1	99
HNF4A	HNF6	USF1	47
HNF4A	CREB1	USF1	134

Figure 4. Real-world example of the *second* type of network structure. A pathway where three regulatory genes are mutually involved in the coordinate regulation of 28-138 genes (adopted from [10]).

³More precisely, in networks 1a and 1b, N = number of all relevant variables, while in network 2, N = number of all relevant variables - 1.

2. Ranking variables by SVM weights. We trained a soft-margin linear SVM classifier [5] for all above generated datasets using SVM penalty parameter $C = \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$. We used libSVM implementation of the SVM algorithm (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>). For each trained classifier, we obtained SVM weights w_i for each variable and ranked variables according the following rule: the larger the weight, the more relevant is the variable.

3. Classifying ranked variables. Once the variables have been ranked based on the training data, we assessed their classification performance on the 5000-sample testing dataset coming from the same experimental setup (network, number of relevant/irrelevant variables, noise level, sample size) using the same value of the SVM penalty parameter C as was used for ranking of variables. Specifically, we classified 10%, 20%, ..., 90%, 100% top-ranked variables. To obtain baselines for classification, we classified the groups of (i) causally relevant variables, (ii) non-causally relevant variables, (iii) all relevant variables, and (iv) irrelevant variables. We used area under ROC curve (AUC) measure to assess classification performance.

4. Assessing ranking of variables. Given SVM weights of variables produced on the training data, we used AUC to analyze how weights discriminate between two groups of variables (e.g., relevant vs. irrelevant). In order to compute AUC, we used group membership (e.g., relevant vs. irrelevant) as the response variable and SVM weights as the predictor.

5. Selecting variables by SVM-RFE. We executed SVM-RFE [7] recursive variable selection algorithm for all generated datasets with ≤ 100 irrelevant variables and training sample size = $\{100, 200, 500\}$ using SVM penalty parameter $C = \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$. To closely follow the published algorithm, we removed one variable during each recursive iteration. We used 75% of the training sample to compute SVM weights of the variable subsets and the remaining 25% to assess their classification performance. The subset of variables returned by SVM-RFE was the one with the highest AUC.

6. Classifying selected variables. Once the variables have been selected based on the training data, we assessed their classification performance on the 5000-sample testing dataset coming from the same experimental setup (network, number of relevant/irrelevant variables, noise level, sample size) using the same value of the SVM penalty parameter C as was used for selection of variables. As previously mentioned, we also assessed classification performance of the baselines.

Results

Below we summarize the main findings of our experiments, while the detailed results are available online from <http://www.dsl-lab.org/supplements/NIPS2006/>. The results presented below illustrate drawbacks of using SVM weight-based methods to discriminate between (i) irrelevant vs. non-causally relevant variables, (ii) non-causally relevant vs. causally relevant variables, and (iii) irrelevant vs. causally relevant variables.

1. SVMs can assign higher weights to the irrelevant variables than to the non-causally relevant ones. Consider results of the simulation experiment with network 1a with 100 relevant and irrelevant variables, training sample size = 100, and no noise. When the SVM-penalty parameter C is small (≤ 0.01), non-causally relevant variables receive higher weights than the irrelevant ones (Figure 5a and Table 1). Surprisingly, when C becomes large (≥ 0.1), we observe the opposite: irrelevant variables receive higher weights than non-causally relevant ones (Figure 5b and Table 1). We note that for both situations, classification performance of the SVMs is excellent and almost the same, thus choosing the best performing classifier cannot help make decisions about relevance of variables ranked by SVM weights (Table 2). On the other hand, as the training sample increases, or the noise level increases, or the number of irrelevant variables becomes larger than the number of relevant variables, the weights of irrelevant variables tend to decrease relatively to the weights of the non-causally relevant variables.

2. SVMs can select irrelevant variables more frequently than the non-causally relevant ones. In general, cases when the relevant variables receive relatively small SVM weights are not novel to the researchers. That is why Guyon et al advocated against ranking variables based on SVM weights, especially when the variables are highly correlated between each other (which is the case for networks 1a and 1b), see section 6.2.1 in [7]. They proposed to use SVM-RFE recursive variable elimination procedure to address this problem.

The results of application of SVM-RFE to the network 1a with 100 relevant and irrelevant variables, training sample size = 100, and no noise are shown in Figure 6. When C is large, irrelevant variables are on average selected more frequently than non-causally relevant ones. The classification performance is not significantly different between small and large C values (Table 3). Again, as the training sample increases, or the noise level increases, or the number of irrelevant variables becomes larger than the number of relevant variables, irrelevant variables tend to be selected less frequently relatively to the non-causally relevant variables.

3. SVMs can assign higher weights to the non-causally relevant variables than to the causally relevant ones.

Now consider results of the simulation experiment with network 2 with 100 relevant and irrelevant variables, training sample size = 500, and no noise. Regardless of the value of C , the non-causally variable Y receives higher weights than the majority of causal ones (Figure 7 and Table 4). We note that as the training sample decreases, the weight of the non-causally relevant variable Y tends to decrease relative to the weights of the causally relevant variables.

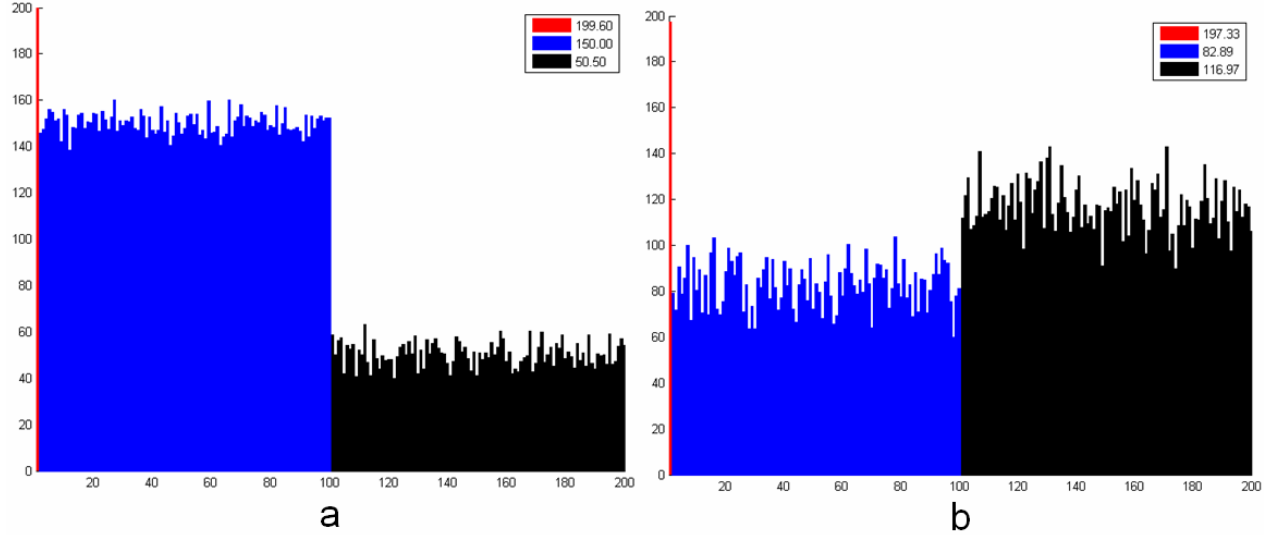


Figure 5. Average ranks of variables (by SVM weights) over 30 random training samples from network 1a. *The larger the rank, the larger is the SVM weight of a variable.* The horizontal axis denotes variables: causally relevant variable X_1 is shown with **red bar** (it is #1 on horizontal axis), non-causally relevant variables X_2, \dots, X_{100} are shown with **blue bars** (#2-#100), irrelevant variables Z_1, \dots, Z_{100} are shown with **black bars** (#101-#200). The vertical axis shows average ranks of the variables. The legend box reports the average rank of a group of variables (causally relevant, non-causally relevant, and irrelevant). Part “a” shows results for $C = 0.001$ and “b” for $C = 1000$.

SVM penalty param C	$N_{\text{sample}} = 100$	$N_{\text{sample}} = 200$	$N_{\text{sample}} = 500$	$N_{\text{sample}} = 1000$
0.001	1.000	1.000	1.000	1.000
0.01	0.834	0.784	0.739	0.811
0.1	0.342	0.406	0.581	0.716
1	0.335	0.423	0.592	0.694
10	0.335	0.423	0.593	0.715
100	0.335	0.423	0.593	0.724
1000	0.335	0.423	0.593	0.726

Table 1. Area under ROC curve (AUC) analysis for discrimination between groups of all relevant and irrelevant variables based on SVM weights (for network 1a). The reported AUC’s in the table are averaged over 30 training samples.

SVM penalty param C	Proportion of top ranked variables by SVM weights									All variables	Causally relevant variable	Non-causally relevant	All relevant variables	Irrelevant variables
	10%	20%	30%	40%	50%	60%	70%	80%	90%					
0.001	0.94	0.933	0.929	0.926	0.924	0.923	0.923	0.922	0.922	0.922	0.955	0.908	0.924	0.499
0.01	0.941	0.936	0.934	0.932	0.931	0.93	0.929	0.928	0.928	0.928	0.955	0.908	0.938	0.499
0.1	0.948	0.934	0.927	0.924	0.922	0.921	0.921	0.921	0.921	0.921	0.955	0.906	0.949	0.498
1	0.93	0.92	0.918	0.916	0.916	0.917	0.917	0.917	0.917	0.917	0.955	0.883	0.94	0.497
10	0.923	0.92	0.918	0.916	0.916	0.917	0.917	0.917	0.917	0.917	0.955	0.851	0.92	0.497
100	0.923	0.92	0.918	0.916	0.916	0.917	0.917	0.917	0.917	0.917	0.955	0.83	0.908	0.497
1000	0.923	0.92	0.918	0.916	0.916	0.917	0.917	0.917	0.917	0.917	0.955	0.83	0.908	0.497

Table 2. Area under ROC curve (AUC) classification performance obtained on the 5,000-sample independent testing set: results for variable ranking based on SVM weights (for network 1a). To account for variance due to small training sample sizes, the reported AUC is the average of AUC’s obtained by classifiers trained on 30 random training samples.

4. SVMs can select *non-causally relevant variables more frequently than the causally relevant ones.* The results of application of SVM-RFE to the network 2 with 100 relevant and irrelevant variables, training sample size = 500, and no noise are shown in Figure 8. When C is large (≥ 0.1), non-causally relevant variable Y is always selected unlike the causally relevant variables. We note that as the training sample decreases, or C decreases, the non-causally relevant variable Y tends to be selected less frequently compared to the causally relevant variables.

5. SVMs can assign higher weights to the *irrelevant variables than to the causally relevant ones.* Consider results of the simulation experiment with network 2 with 100 relevant and irrelevant variables, training sample size = 100, and no noise. Regardless of the value of C , there are some irrelevant variables that receive higher weights than the majority of causally relevant ones (see variables #155 and #185 in Figure 9 and Table 5). However, as the training sample increases, the weights of irrelevant variables tend to decrease relative to the weights of the causally relevant variables.

6. SVMs can select *irrelevant variables more frequently than the causally relevant ones.* The results of application of SVM-RFE to the network 2 with 100 relevant and irrelevant variables, training sample size = 100, and no noise are provided in Figure 10. Regardless of the value of C , a few irrelevant variables are selected more frequently than the majority of causally relevant ones. In fact, variable #190 shown in Figure 10 is selected by SVM-RFE more frequently than 90 out of 100 causally relevant variables! As the training sample increases, the irrelevant variables tend to be selected less frequently compared to the causally relevant variables. However, in the examples that we considered with training sample up to 500 there still were some irrelevant variables that were selected more frequently than the majority of causally relevant ones.

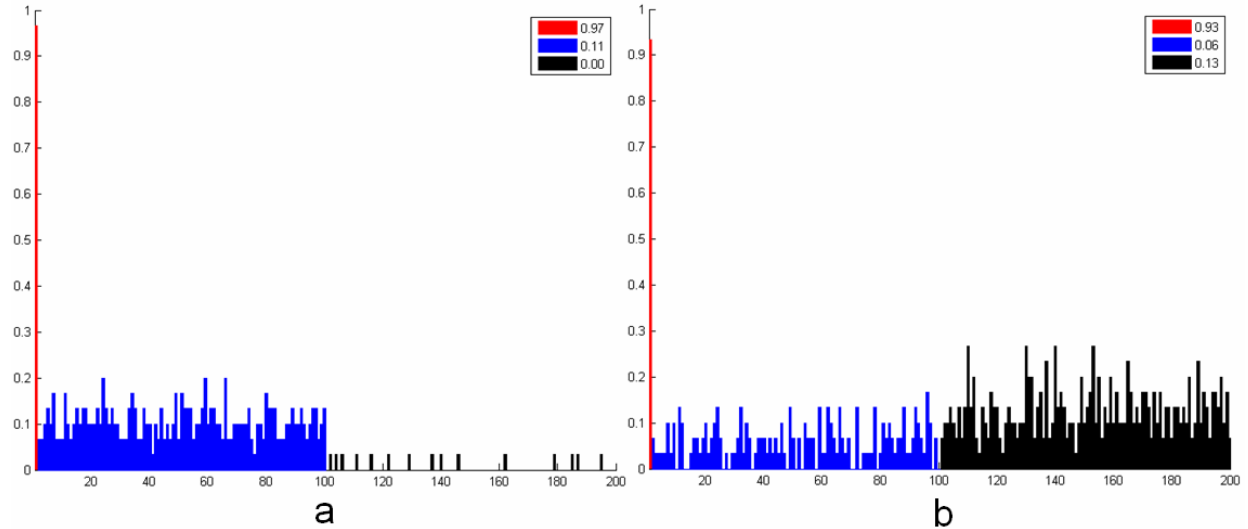


Figure 6. Probability of selecting variables (by SVM-RFE) estimated over 30 random training samples from network 1a. The horizontal axis denotes variables: causally relevant variable X_1 is shown with **red bar** (it is #1 on horizontal axis), non-causally relevant variables X_2, \dots, X_{100} are shown with **blue bars** (#2-#100), irrelevant variables Z_1, \dots, Z_{100} are shown with **black bars** (#101-#200). The vertical axis shows probabilities of selecting variables. The legend box reports the average probability of selecting a variable in a group (causally relevant, non-causally relevant, and irrelevant). Part “a” shows results for $C = 0.001$ and “b” for $C = 1000$.

SVM penalty param C	Selected variables by SVM-RFE	Causally relevant variable	Non-causally relevant variables	All relevant variables	Irrelevant variables
0.001	0.948	0.955	0.908	0.924	0.499
0.01	0.948	0.955	0.908	0.938	0.499
0.1	0.949	0.955	0.906	0.949	0.498
1	0.942	0.955	0.883	0.94	0.497
10	0.937	0.955	0.851	0.92	0.497
100	0.94	0.955	0.83	0.908	0.497
1000	0.933	0.955	0.83	0.908	0.497

Table 3. Area under ROC curve (AUC) classification performance obtained on the 5,000-sample independent testing set: results for variable selection by SVM-RFE (for network 1a). To account for variance due to small training sample sizes, the reported AUC is the average of AUC’s obtained by classifiers trained on 30 random training samples.

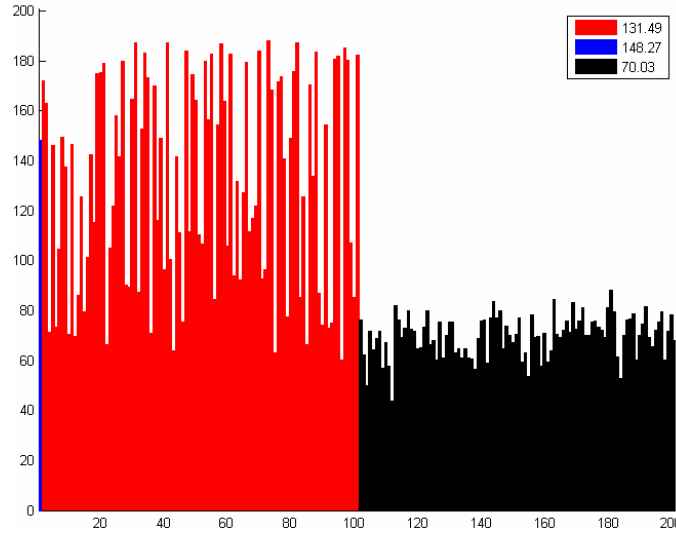


Figure 7. Average ranks of variables (by SVM weights) over 30 random training samples from network 2. *The larger the rank, the larger is the SVM weight of a variable.* The horizontal axis denotes variables: causally relevant variables X_1, \dots, X_{100} are shown with **red bars** (#2-#101 on horizontal axis), non-causally relevant variable Y is shown with **blue bar** (#1), irrelevant variables Z_1, \dots, Z_{100} are shown with **black bars** (#102-#201). The vertical axis shows average ranks of the variables. The legend box reports the average rank of a group of variables (causally relevant, non-causally relevant, and irrelevant). The results are shown for $C = 1$.

SVM penalty param C	$N_{\text{sample}} = 100$	$N_{\text{sample}} = 200$	$N_{\text{sample}} = 500$	$N_{\text{sample}} = 1000$
0.001	0.604	0.535	0.502	0.49
0.01	0.602	0.536	0.502	0.482
0.1	0.61	0.531	0.489	0.484
1	0.611	0.528	0.498	0.485
10	0.611	0.528	0.498	0.484
100	0.611	0.528	0.498	0.484
1000	0.611	0.528	0.498	0.484

Table 4. Area under ROC curve (AUC) analysis for discrimination between groups of causally relevant and non-causally relevant variables based on SVM weights (for network 2). The reported AUC's in the table are averaged over 30 training samples.

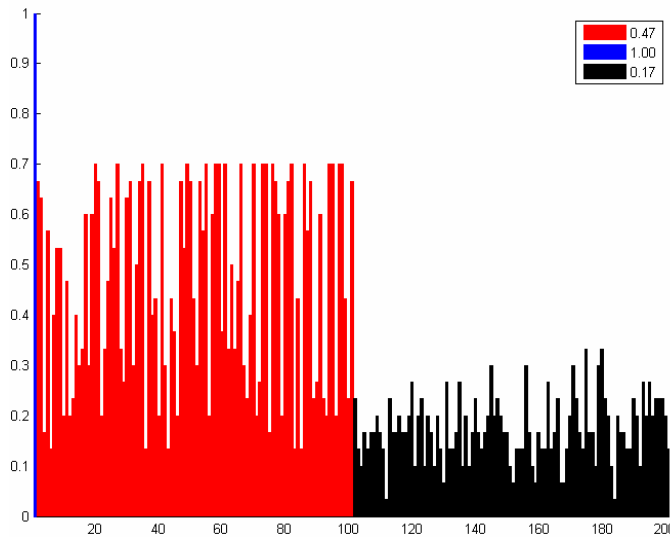


Figure 8. Probability of selecting variables (by SVM-RFE) estimated over 30 random training samples from network 2. The horizontal axis denotes variables: causally relevant variables X_1, \dots, X_{100} are shown with **red bars** (#2-#101 on horizontal axis), non-causally relevant variable Y is shown with **blue bar** (#1), irrelevant variables Z_1, \dots, Z_{100} are shown with **black bars** (#102-#201). The vertical axis shows probabilities of selecting variables. The legend box reports the average probability of selecting a variable in a group (causally relevant, non-causally relevant, and irrelevant). The results are shown for $C = 1$.

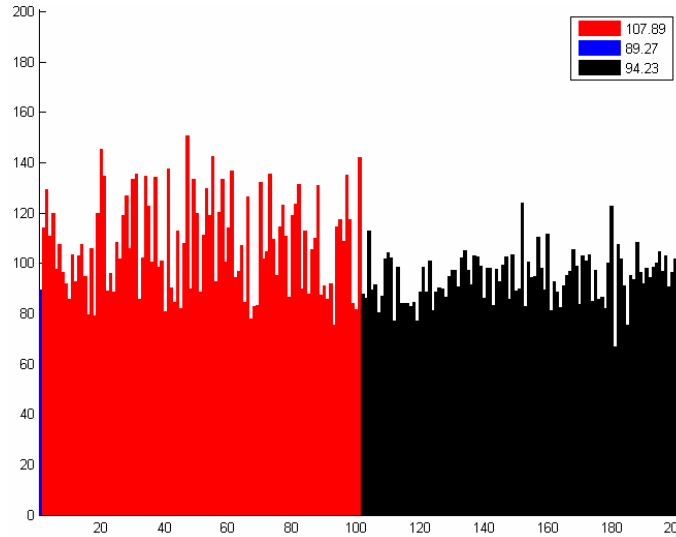


Figure 9. Average ranks of variables (by SVM weights) over 30 random training samples from network 2. *The larger the rank, the larger is the SVM weight of a variable.* The horizontal axis denotes variables: causally relevant variables X_1, \dots, X_{100} are shown with **red bars** (#2-#101 on horizontal axis), non-causally relevant variable Y is shown with **blue bar** (#1), irrelevant variables Z_1, \dots, Z_{100} are shown with **black bars** (#102-#201). The vertical axis shows average ranks of the variables. The legend box reports the average rank of a group of variables (causally relevant, non-causally relevant, and irrelevant). The results are shown for $C = 0.001$.

SVM penalty param C	$N_{\text{sample}} = 100$	$N_{\text{sample}} = 200$	$N_{\text{sample}} = 500$	$N_{\text{sample}} = 1000$
0.001	0.565	0.65	0.737	0.799
0.01	0.571	0.654	0.756	0.836
0.1	0.58	0.664	0.803	0.878
1	0.58	0.662	0.805	0.891
10	0.58	0.662	0.805	0.893
100	0.58	0.662	0.805	0.893
1000	0.58	0.662	0.805	0.893

Table 5. Area under ROC curve (AUC) analysis for discrimination between groups of causally relevant and irrelevant variables based on SVM weights (for network 2). The reported AUC's in the table are averaged over 30 training samples.

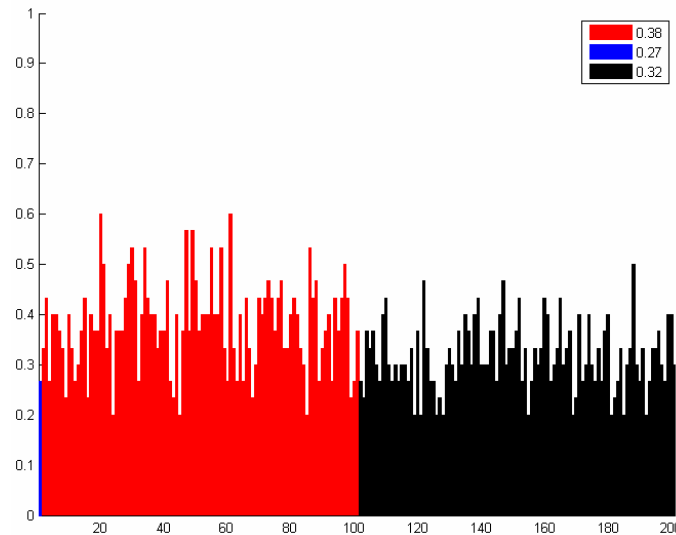


Figure 10. Probability of selecting variables (by SVM-RFE) estimated over 30 random training samples from network 2. The horizontal axis denotes variables: causally relevant variables X_1, \dots, X_{100} are shown with **red bars** (#2-#101 on horizontal axis), non-causally relevant variable Y is shown with **blue bar** (#1), irrelevant variables Z_1, \dots, Z_{100} are shown with **black bars** (#102-#201). The vertical axis shows probabilities of selecting variables. The legend box reports the average probability of selecting a variable in a group (causally relevant, non-causally relevant, and irrelevant). The results are shown for $C = 0.001$.

Theoretical Examples

Any algorithm that explicitly or implicitly makes determinations about local causality that are inconsistent with the causal process that generated the data will also make incorrect determinations about the complete causal process. Hence local causal consistency (to the generating causal process) is required for global consistency. This preamble is necessary because variable selection algorithms are not designed to return full causal graphs, but they are often used to select candidates for local causes and effects.

Example 1. First, we provide an example where SVMs assign the largest weight to a non-causally relevant variable. Consider the network structure of type 2 as shown in Figure 2, where:

- X_1 and X_2 are binary with $P(X_1=-1) = 1/2$, $P(X_1=1) = 1/2$, $P(X_2=-1) = 1/2$, and $P(X_2=1) = 1/2$.
- Y is a “synthesis variable” with the following function: $Y = \frac{X_1 + X_2}{\sqrt{2}}$.
- There are no irrelevant variables.
- T is a binary response variable defined as $T = \text{sign}(X_1 + X_2 - 1)$.

We note that variables X_1 , X_2 , and Y have expected value 0 and variance 1. The application of linear SVMs results in the following weights: $1/2$ for X_1 , $1/2$ for X_2 , and $1/\sqrt{2}$ for Y . Therefore, the non-causally relevant variable Y receives higher SVM weight than any causally relevant one (X_1 or X_2).

Example 2. In this example we show that a fundamental weakness of the maximum-gap inductive bias, as employed in SVMs, is its local causal inconsistency. Consider a scenario (Figure 11) where we wish to discover the direct causes of a response variable T , from observations about variables X , Y , T . Assume for simplicity that T is a terminal variable and thus X and Y precede it in time. For example, T can be a clinical phenotype and X, Y can be gene expression values. The causal process that generates the data is seen in the upper right corner of Figure 11.

As can be seen in the left part of the Figure 11, the SVM classifier can perfectly predict T using X and Y as predictors. In doing so it prefers the classifier with gap $G1$ to the classifier with smaller gap $G2$. The preferred classifier assigns non-zero (and in fact equal) weights to both X , Y thereby admitting Y in the local causal neighborhood if selected variables are interpreted causally. However, X renders Y independent from T and not vice versa. More generally in distributions where the Causal Markov Condition holds (which states that a variable is independent from all its non-effects given its direct causes) SVMs will occasionally fail to detect that Y is not a local cause of T . State-of-the-art causal discovery algorithms do not face this problem, however [12].

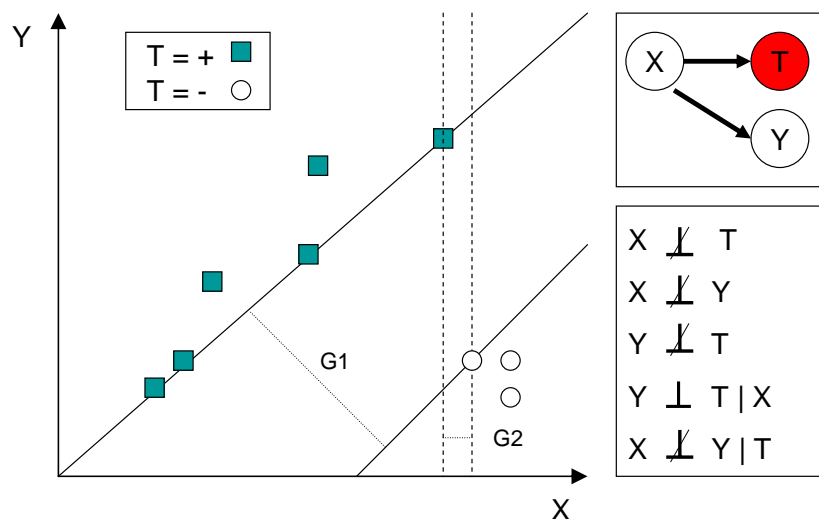


Figure 11. The maximum-gap inductive bias is inconsistent with local causal discovery. The symbol \perp means independent, and $\not\perp$ means dependent.

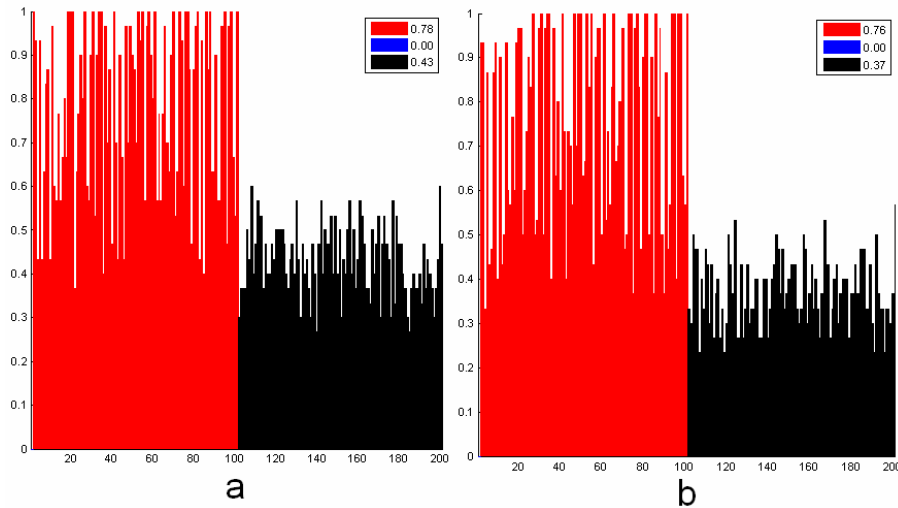


Figure 12. Probability of selecting variables (by polynomial SVM-RFE) estimated over 30 random training samples from network 2. The horizontal axis denotes variables: causally relevant variables X_1, \dots, X_{100} are shown with **red bars** (#2-#101 on horizontal axis), non-causally relevant variable Y is shown with **blue bar** (#1), irrelevant variables Z_1, \dots, Z_{100} are shown with **black bars** (#102-#201). The vertical axis shows probabilities of selecting variables. The legend box reports the average probability of selecting a variable in a group (causally relevant, non-causally relevant, and irrelevant). Part “a” shows results for polynomial kernel degree = 2 and “b” for degree = 3.

Discussion

The above experiments and examples were limited to the linear SVM weight-based methods. An interesting direction for further research is the analysis of nonlinear SVM techniques. We have conducted preliminary experiments to study the polynomial SVM-RFE (see section 6.3 in [7]) in the simulation example where the non-causally relevant variable was selected more frequently than any causally relevant one by SVM-RFE (see section 4 in the Results section). The results are shown in Figure 12 for the same experimental parameters and C value as used in Figure 8. Surprisingly, the polynomial SVM-RFE (both for kernel of degree 2 and 3) never selected the non-causally relevant variable Y . On the other hand, both SVM-RFE and polynomial SVM-RFE resulted in classifiers with excellent and statistically identical performances (0.923 AUC for SVM-RFE vs. 0.936 AUC and 0.933 AUC for polynomial SVM-RFE for kernel degree 2 and 3, respectively). Following the structural risk minimization principle [16], an analyst will select the simplest best performing model which yields the wrong causal interpretation in this problem. It is also worthwhile to mention that the polynomial SVM-RFE selects some irrelevant variables more frequently than the majority of causally relevant ones (see variables in the region #195-#200 in Figure 12).

The simulations and theoretical examples presented in this paper suggest that the SVM weight-based methods studied cannot readily uncover causal relations even in simple and non-contrived causal processes. On the other hand, the framework of formal causal discovery [12] provides algorithms that can solve these problems, for example [4,14,15]. In a forthcoming extended paper, we plan to apply causal discovery algorithms to these problems and compare with SVM weight-based methods.

Likewise, we also plan to study methods based on modified SVM formulations, such as methods with 0-norm and 1-norm penalties [17,18]. These techniques typically result in fewer selected variables, but may sacrifice predictive performance compared to the standard SVMs with 2-norm weight penalty.

Finally, another opportunity for further research is to extend the empirical evaluation to different distributions.

Conclusions

The main conclusion of this study is that *causal interpretation of current SVM weight-based variable selection techniques must be conducted with great caution by practitioners*. In addition, we provided examples where SVMs assign higher weights or select (in the context of SVM-RFE) irrelevant variables more frequently than the relevant ones. Finally, we provided theoretical examples to explain why non-causally relevant variables may be preferred by SVMs compared to the causally relevant ones. In particular, we showed that the inductive bias employed by SVMs is locally causally inconsistent. New SVM methods may be needed to address this issue and this is an exciting and challenging area of research.

Acknowledgements

This research was supported by NIH grant RO1 LM007948–01. The authors would like to thank Efi Kokkotou and Siddharth Pratap.

Reference List

1. Aliferis CF, Statnikov A, Kokkotou E, Massion PP, Tsamardinos I: **Local regulatory-network inducing algorithms for biomarker discovery from mass-throughput datasets.** (submitted) *Technical Report DSL 06-05* 2006.
2. Aliferis CF, Statnikov A, Massion PP: **Pathway induction and high-fidelity simulation for molecular signature and biomarker discovery in lung cancer using microarray gene expression data.** *Proceedings of the 2006 American Physiological Society Conference "Physiological Genomics and Proteomics of Lung Disease"* 2006.
3. Aliferis CF, Statnikov A, Tsamardinos I, Kokkotou E, Massion PP: **Application and comparative evaluation of causal and non-causal feature selection algorithms for biomarker discovery in high-throughput biomedical datasets.** *Proceedings of the NIPS 2006 Workshop on Causality and Feature Selection* 2006.
4. Aliferis CF, Tsamardinos I, Statnikov A: **HITON: a novel Markov blanket algorithm for optimal variable selection.** *AMIA 2003 Annual Symposium Proceedings* 2003, 21-25.
5. Boser BE, Guyon IM, Vapnik VN: **A training algorithm for optimal margin classifiers.** *Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT)* 1992, 144-152.
6. Guyon I, Elisseeff A: **An introduction to variable and feature selection.** *Journal of Machine Learning Research* 2003, **3**: 1157-1182.
7. Guyon I, Weston J, Barnhill S, Vapnik V: **Gene selection for cancer classification using support vector machines.** *Machine Learning* 2002, **46**: 389-422.
8. Hardin D, Tsamardinos I, Aliferis CF: **A theoretical characterization of linear SVM-based feature selection.** *Proceedings of the Twentieth First International Conference on Machine Learning (ICML)* 2004.
9. Kohavi R, John GH: **Wrappers for feature subset selection.** *Artificial Intelligence* 1997, **97**: 273-324.
10. Odom DT, Dowell RD, Jacobsen ES, Nekludova L, Rolfe PA, Danford TW *et al.*: **Core transcriptional regulatory circuitry in human hepatocytes.** *Mol Syst Biol* 2006, **2**: 2006.
11. Rakotomamonjy A: **Variable selection using SVM-based criteria.** *Journal of Machine Learning Research* 2003, **3**: 1357-1370.
12. Spirtes P, Glymour CN, Scheines R: *Causation, prediction, and search*, 2nd edn. Cambridge, Mass: MIT Press; 2000.
13. Tsamardinos I, Aliferis CF: **Towards principled feature selection: relevancy, filters and wrappers.** *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics (AI & Stats)* 2003.
14. Tsamardinos I, Aliferis CF, Statnikov A: **Time and sample efficient discovery of Markov blankets and direct causal relations.** *Proceedings of the Ninth International Conference on Knowledge Discovery and Data Mining (KDD)* 2003, 673-678.
15. Tsamardinos I, Brown LE, Aliferis CF: **The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm.** *Machine Learning* 2006, **65**: 31-78.
16. Vapnik VN: *Statistical learning theory*. New York: Wiley; 1998.
17. Weston J, Elisseeff A, Scholkopf B, Tipping M: **Use of the zero-norm with linear models and kernel methods.** *Journal of Machine Learning Research* 2003, **3**: 1439-1461.
18. Zhu J, Rosset S, Hastie T, Tibshirani R: **1-norm support vector machines.** *Advances in Neural Information Processing Systems (NIPS)* 2004, **16**.