Eric V. Strobl* and Shyam Visweswaran

# Markov Boundary Discovery with Ridge Regularized Linear Models

DOI 10.1515/jci-2015-0011

**Abstract:** Ridge regularized linear models (RRLMs), such as ridge regression and the SVM, are a popular group of methods that are used in conjunction with coefficient hypothesis testing to discover explanatory variables with a significant multivariate association to a response. However, many investigators are reluctant to draw causal interpretations of the selected variables due to the incomplete knowledge of the capabilities of RRLMs in causal inference. Under reasonable assumptions, we show that a modified form of RRLMs can get "very close" to identifying a subset of the Markov boundary by providing a worst-case bound on the space of possible solutions. The results hold for any convex loss, even when the underlying functional relationship is nonlinear, and the solution is not unique. Our approach combines ideas in Markov boundary and sufficient dimension reduction theory. Experimental results show that the modified RRLMs are competitive against state-of-the-art algorithms in discovering part of the Markov boundary from gene expression data.

**Keywords:** Markov boundary, ridge regularization, linear models

## 1 Introduction

Inferring causal relationships between random variables from observational data is a difficult task. However, algorithms that can accurately identify causal relations may significantly speed up scientific investigations. In many applications, scientists are primarily interested in recovering a restricted group of variables which share causal relationships with a given response variable. For example, biologists may be interested in using gene expression levels to identify new targets which effect cancer survivability. Algorithms which can discover variables that share causal relationships with the response can thus be very useful in practice.

Under mild assumptions (details in Section 2.3), the Markov boundary is a unique group of random variables located within the close vicinity of the response variable in a casual directed graph, where causal relations are modeled by directed edges between variables. In a causal directed acyclic graph, the Markov boundary consists of the response's direct causes, direct effects, and direct causes of the direct effects [1, 2]. Otherwise, the Markov boundary can be a subset of those variables [1, 3]. Note that certain manipulations of the direct causes may modify the distribution of the response, and certain manipulations of the direct effects or direct causes of the direct effects may modify the distributions of the response's consequences. The Markov boundary can thus be a small set of variables that share important causal relationships with a variable of interest.

Variables in the Markov boundary exhibit the following conditional independence relation with other variables in the dataset:

$$Y \perp (\boldsymbol{X} \backslash \boldsymbol{M}) | \boldsymbol{M}, \tag{1}$$

where $Y$ denotes the response variable, $\boldsymbol{X}$ all variables other than $Y$, $\boldsymbol{M}$ the Markov boundary, and $\perp$ statistical independence. Unlike the Markov blanket, the Markov boundary is defined so that no proper

***Corresponding author: Eric V. Strobl,** Center for Causal Discovery, Department of Biomedical Informatics, University of Pittsburgh School of Medicine, 5607 Baum Boulevard, Pittsburgh, PA 15206, USA, E-mail: evs17@pitt.edu
**Shyam Visweswaran,** Center for Causal Discovery, Department of Biomedical Informatics, University of Pittsburgh School of Medicine, 5607 Baum Boulevard, Pittsburgh, PA 15206, USA, E-mail: shv3@pitt.edu

subset of $\boldsymbol{M}$ also satisfies eq. (1) [4]. This fact seems to suggest that Markov boundary discovery is closely related to the variable selection problem. This intuition is indeed true (details in Sections 2.4 and 3.2), but most variable selection methods unfortunately do not attempt to recover the Markov boundary. Their primary goal is to identify a set of variables that maximizes the predictive accuracy of a model, so the causal interpretation of the selected variables has often been heuristic in nature. This is unfortunate because there is often a strong desire to interpret the selected variables as having some sort of causal relationship with the response. Whole research programs have in fact been built on experiments motivated by causal hypotheses that were generated by non-causal variable selection algorithms [5–8]. It is thus important to develop a theoretical understanding of both the advantages and disadvantages of using particular selection strategies for causal inference.

Ridge regularized linear models (RRLMs), such as ridge regression and the support vector machine (SVM),[1] are popular methods that are used to identify variables with a multivariate association with the response. To understand ridge or $L^2$-regularization intuitively, recall that the mean squared error (MSE) of the estimator $\hat{\beta}$ with respect to the variable coefficients $\beta \in \mathbb{R}^p$ can be decomposed into a bias term and a variance term. Ridge regularization penalizes the size of the variable coefficients by shrinking the values of the estimate $\hat{\beta}$ towards zero; this introduces bias but reduces the variance of the estimate. The method is thus particularly useful in the high dimensional low sample size (HDLSS) context, where the variance of $\hat{\beta}$ can be large. In fact, in ridge regression, there always exists a regularization constant $\lambda \in \mathbb{R}^+$ such that the MSE of $\hat{\beta}$ estimated via ridge regression is smaller than the MSE of $\hat{\beta}$ estimated via ordinary least squares [9]. Moreover, in the context of linear SVMs, recall that the distance between the support vectors is $2/\|\beta\|_2$ in the separable binary case. As a result, minimizing $\|\beta_2\|_2$ is equivalent to maximizing the margin distance, an important quantity which can be used to bound the SVM's generalization error [10].

RRLMs are also popular for several other reasons besides variance reduction and margin maximization. First, most RRLMs can be optimized efficiently (e.g., Glmnet [11] and LIBSVM [12]). As a result, permutation analyses can be employed to recover p-values for each variable; these p-values are often the primary quantity of interest in both variable selection and ranking, since they *may* help quantify the probability that subsequently conducted experiments will fail to identify an effect. RRLMs are also used in the HDLSS context, where detection of nonlinearities can be difficult, and linear models are known to often predict better than or comparably with nonlinear methods. Finally, variable importance can be easily assessed in linear models using the absolute values of the coefficients, whereas complex strategies may be needed to recover variable importance values from nonlinear methods. This implies that backward elimination can also be carried out efficiently with RRLMs by iteratively eliminating the variable associated with the coefficient with the smallest absolute value. Together, these strengths highlight the importance of developing an understanding of the causal inference capabilities of RRLMs.

In this paper, we characterize the ability of RRLMs in identifying the Markov boundary. Our results show that a modified form of RRLMs can get "very close" to identifying a subset of the Markov boundary by providing a worst-case bound on the space of possible solutions. This result holds even if the underlying functional relationship between the explanatory variables and response variable is nonlinear. Furthermore, the solution to each RRLM's optimization problem does not need to be unique. The theory is based on the fact that the expectation of the explanatory variables can almost always be written as a linear function of their lower dimensional projections $\eta^T \boldsymbol{X}$ in the high dimensional context [13]. We also present experimental results which demonstrate that the modified RRLMs are competitive when compared with more sophisticated methods for discovering a subset of the Markov boundary from HDLSS gene expression data.

[1] The linear SVM can be implemented with $L^1$ and/or $L^2$-regularization. In this paper, we refer to the original and most popular version of the SVM with $L^2$-regularization and the hinge loss [40, 41, 42].

# 2 Background

## 2.1 Notation

Unless specified otherwise, upper case letters in italics denote random variables (e.g., $A$, $B$, $C$) and upper case bold letters in italics denote random variable sets or random column vectors ($\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$). We reserve $Y$ for the response variable, and $\boldsymbol{X}$ for the set or column vector of all $p$ explanatory variables. We also use the notation $\boldsymbol{A} \subseteq \boldsymbol{B}$ to mean $\boldsymbol{A}$ is a subset of $\boldsymbol{B}$, and $\boldsymbol{A} \subset \boldsymbol{B}$ to mean $\boldsymbol{A}$ is a proper subset of $\boldsymbol{B}$.

We will be arranging both vectors and matrices in this paper, so we use the following notation to distinguish between operations involving rows versus columns. If $\boldsymbol{A}$ is a $a \times 1$ vector and $\boldsymbol{B}$ is a $b \times 1$ vector, then we use the notation ($\boldsymbol{A}$; $\boldsymbol{B}$) to denote that $\boldsymbol{B}$ is appended to the rows of $\boldsymbol{A}$ to create a $(a+\mathrm{b}) \times 1$ column vector. On the other hand, we use the notation ($\boldsymbol{A}$, $\boldsymbol{B}$) to denote that $\boldsymbol{B}$ is appended to the columns of $\boldsymbol{A}$ to create a $a \times 2$ matrix when $a=b$.

A graph consists of nodes and edges, where nodes are random variables and directed edges are causal relationships between two variables. If a graph contains a directed edge between two variables $A$ and $B$ such that $A \rightarrow B$, then $A$ is a direct cause or parent of $B$, and $B$ is a direct effect or child of $A$. A node $A$ is a direct cause of the direct effect or a spouse of $B$, if both $A$ and $B$ share a common child node.

We write $\boldsymbol{A} \perp \boldsymbol{B} | \boldsymbol{C}$ when two sets of variables $\boldsymbol{A}$ and $\boldsymbol{B}$ are conditionally independent given a third set of variables $\boldsymbol{C}$. We also write $\boldsymbol{A} \perp \boldsymbol{B}$ when the conditioning set is empty to denote that $\boldsymbol{A}$ and $\boldsymbol{B}$ are independent.

## 2.2 Distribution Theory

The following is an important property of a probability distribution, which we will often refer to in this paper:

**Definition 2.2.1. (Intersection Property)** Let $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$ be any four subsets of potentially overlapping variables from ($Y$; $\boldsymbol{X}$) with joint probability distribution $\mathbb{P}$. The distribution $\mathbb{P}$ is said to satisfy the intersection property if $\boldsymbol{A} \perp \boldsymbol{B} | (\boldsymbol{C} \cup \boldsymbol{D})$ and $\boldsymbol{A} \perp \boldsymbol{D} | (\boldsymbol{C} \cup \boldsymbol{B}) \Rightarrow \boldsymbol{A} \perp (\boldsymbol{B} \cup \boldsymbol{D}) | \boldsymbol{C}$.

There are two sufficient conditions for the intersection property:

**Proposition 2.2.2.** If $\mathbb{P}$ has a strictly positive density or has a continuous density with path-connected support, then $\mathbb{P}$ satisfies the intersection property.

By continuous density with path-connected support, we mean that an absolutely continuous distribution has a continuous density whose support can be connected by axis-parallel lines; this is in fact both a necessary and sufficient condition. Note that we will always assume that a distribution has a density in this paper, so we will use the words "distribution" and "density" interchangeably. For the proof of Proposition 2.2.2, the part regarding strict positivity is provided in [3, 14], and the part regarding path-connected support is provided in [15].

We believe that these sufficient conditions are mild and are satisfied in most distributions encountered in practice. Strict positivity is considered reasonable whenever there is uncertainty about the data [3]. Moreover, it is reasonable to assume that almost all distributions with continuous densities have path-connected support in biology. We nonetheless must acknowledge that several investigators have challenged these assumptions (e.g. [4, 16, 17]). We do not claim that strict positivity or path-connected support holds in all biological datasets; however, we believe that the empirical results used to refute these assumptions are still inconclusive, since some of the same investigators studying HDLSS molecular data have warned that their experimental results can also hold when sample sizes are insufficient [4, 18–20].

## 2.3 Markov boundary theory

We now define a special group of variables which helps characterize a conditional independence relation concerning the response variable $Y$:

**Definition 2.3.1: Markov blanket.** A Markov blanket $\boldsymbol{M}$ of a response variable $Y$ in the joint probability distribution $\mathbb{P}$ over variables $\boldsymbol{X}$ is a set of variables conditioned on which all other variables are independent of $Y$; that is, for every $\boldsymbol{F} \subseteq (\boldsymbol{X} \backslash \boldsymbol{M}), Y \perp \boldsymbol{F} | \boldsymbol{M}$.

The Markov blanket of $Y$ is shown to exhibit the following relation using eq. (1):

**Proposition 2.3.2.** We have $Y \perp (\boldsymbol{X} \backslash \boldsymbol{M}) | \boldsymbol{M} \Leftrightarrow Y \perp \boldsymbol{X} | \boldsymbol{M}$.

The proof follows directly from Proposition 4.4 of [21]. Note that a trivial Markov blanket of $Y$ is $\boldsymbol{X}$. As a result, we more specifically define a minimal Markov blanket:

**Definition 2.3.3: Markov boundary.** If no proper subset of $\boldsymbol{M}$ satisfies the definition of a Markov blanket of $Y$, then $\boldsymbol{M}$ is the Markov boundary of $Y$.

We now have a sufficient condition for the uniqueness of the Markov boundary of $Y$:

**Theorem 2.3.4** [3]. If a joint probability distribution $\mathbb{P}$ over variables ($Y$; $\boldsymbol{X}$) satisfies the intersection property, then for each $V \in (Y;X)$, there exists a unique Markov boundary of $V$.

Note that some investigators have studied violations of the intersection property, such as in deterministic relations, where uniqueness of the Markov boundary may not hold [4, 16]. However, we assume that the intersection property holds in this paper.

We define an important condition which we will use in the next theorem:

**Definition 2.3.5: Global Markov condition** [22, 23]. The joint probability distribution $\mathbb{P}$ over variables $\boldsymbol{H}$ satisfies the global Markov condition for a directed graph $\mathbb{G} = <\boldsymbol{H}, \mathbb{E}>$ if and only if for any three disjoint subsets of variables $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ from $\boldsymbol{H}$, if $\boldsymbol{A}$ is d-separated from $\boldsymbol{B}$ given $\boldsymbol{C}$ in $\mathbb{G}$ then $\boldsymbol{A}$ is conditionally independent of $\boldsymbol{B}$ given $\boldsymbol{C}$ in $\mathbb{P}$.

The following theorem guarantees that the Markov boundary is a subset of the children, parents and spouses of the response variable, if the global Markov condition is satisfied:

**Theorem 2.3.6** [1, 3]**.** If a joint probability distribution $\mathbb{P}$ satisfies the global Markov condition for directed graph $\mathbb{G}$, then the set of children, parents and spouses of $Y$ is a Markov blanket of $Y$.

## 2.4 The Markov blanket and variable selection

This section connects the ideas of a Markov blanket to generic variable selection. We first consider an optimal set of variables for predicting $Y$.

**Definition 2.4.1: Optimal predictor** [4]. Given a dataset $\mathbb{D}$ (a sample from distribution $\mathbb{P}$) for variables ($Y$; $\boldsymbol{X}$), a learning algorithm $\mathbb{L}$, and a performance metric $\mathbb{M}$ to assess the learner's models, a variable set $\boldsymbol{V} \subseteq \boldsymbol{X}$ is an optimal predictor of $Y$ if $\boldsymbol{V}$ maximizes $\mathbb{M}$ for predicting $Y$ using learner $\mathbb{L}$ in the dataset $\mathbb{D}$.

Examples of performance metrics include the log-likelihood, negative mean squared error, and negative hinge loss. We can equivalently consider minimizing a loss functional instead of maximizing a performance metric. The following theorem provides necessary and sufficient conditions for the Markov blanket to be an optimal predictor.

**Theorem 2.4.2** [4]**.** If $\mathbb{M}$ is a performance metric that is maximized only when the conditional probability distribution $\mathbb{P}(Y|\boldsymbol{X})$ is estimated accurately, and $\mathbb{L}$ is a learning algorithm that can approximate any conditional probability distribution, then $\boldsymbol{M}$ is a Markov blanket of $Y$ if and only if it is an optimal predictor of $Y$.

## 2.5 Sufficient dimension reduction theory

Sufficient dimension reduction (SDR) can be viewed as a generalization of Markov boundary discovery. SDR attempts to find a matrix $\eta \in \mathbb{R}^{p \times d}$, where $d \leq p$ such that:

$$Y \perp \boldsymbol{X} | \eta^T \boldsymbol{X}. \tag{2}$$

Notice that $\eta^T \boldsymbol{X}$ is a lower dimensional linear combination of $\boldsymbol{X}$. In contrast, Markov boundary discovery algorithms first consider $\eta = I_p$, and then use combinatorial optimization methods (e.g., constraint-based search) to set some of the diagonal elements to zero. We will see that the added flexibility of considering any lower dimensional linear combination of $\boldsymbol{X}$ which satisfies eq. (2) will allow us to convert the traditional combinatorial optimization approach of Markov boundary discovery to an easier, continuous optimization problem without sacrificing fidelity.

We provide the following two definitions from [21]. Note that we do not restrict ourselves to a continuous response or regression. We will use the notation $Y|\boldsymbol{X}$ to denote the response variable $Y$ given the explanatory variables $\boldsymbol{X}$; one can then speak of, for example, the probability distribution or expectation of $Y$ given $\boldsymbol{X}$, which we write as $\mathbb{P}(Y|\boldsymbol{X})$ and $E(Y|\boldsymbol{X})$, respectively. For instance, we may have:

$$Y|\boldsymbol{X} = X_1 + 2X_2 + \varepsilon,$$

where $\varepsilon$ is Gaussian noise and $\varepsilon \perp \boldsymbol{X}$.

**Definition 2.5.1: Dimension reduction subspace.** If $Y \perp \boldsymbol{X} | \eta^T \boldsymbol{X}$, then the column space of $\eta$ is called a dimension reduction subspace (DRS) for $Y|\boldsymbol{X}$.

The column space of $\eta$ refers to the space spanned by the columns of $\eta$ such that the columns form a basis. We will denote the column space of $\eta$ as $S(\eta)$. Also:

**Definition 2.5.2: Central dimension reduction subspace.** A subspace $S_{Y|\boldsymbol{X}}$ is a central DRS for $Y \mid \boldsymbol{X}$ if $S_{Y|\boldsymbol{X}}$ is a DRS and $S_{Y|\boldsymbol{X}} \subseteq S_{DRS}$ for all DRSs $S_{DRS}$.

The following is an important sufficient condition for the existence of a central DRS:

**Theorem 2.5.3** [21]**.** Let $S(\alpha)$ and $S(\phi)$ be DRSs for $Y \mid \boldsymbol{X}$. If $\boldsymbol{X}$ has a density $f(a) > 0$ for $a \in \Omega_{\boldsymbol{X}} \subseteq \mathbb{R}^p$ and $f(a) = 0$ otherwise, and if $\Omega_X$ is a convex set, then $S(\alpha) \cap S(\phi)$ is a DRS.

The central DRS for $Y|\boldsymbol{X}$ thus exists if the density of $\boldsymbol{X}$ has convex support because the intersection of all DRSs is the central DRS, $\cap S_{DRS} = S_{Y|\boldsymbol{X}}$. A similar result also exists for discrete distributions with connected support (see Problem 6.5 in [21]).

# 3 Theoretical results

## 3.1 Summary

We provide a summary of the results of this section. Consider the following assumptions:

(1) The global Markov condition holds.

(2) The joint probability distribution of $(Y; \boldsymbol{X})$ satisfies the linear intersection property (defined in Section 3.3).

(3) $\sum_{\boldsymbol{X}}$ is positive definite.

(4) $E(\boldsymbol{X}|\eta^T\boldsymbol{X})$ is a linear function of $\eta^T\boldsymbol{X}$ when $Y \perp \boldsymbol{X}|\eta^T\boldsymbol{X}$.

(5) The matrix $\beta^\star$ is a non-zero matrix, and the solution to the following optimization problem:

$$\arg\min_{\alpha,\beta} E\{u(\alpha+\beta^T\boldsymbol{X}, Y)\} + \lambda tr\left(\beta^T \sum\nolimits_{\boldsymbol{X}} \beta\right).$$

where $\alpha \in \mathbb{R}^K$, $\beta \in \mathbb{R}^{p\times K}$, $u(\cdot,\cdot)$ is any convex functional, $tr(\ \cdot)$ denotes the trace, $\lambda \in \mathbb{R}^+$, and $K \in \mathbb{Z}^+$. Notice that a covariance matrix $\sum_X$ has been added into the ridge regularization, and the regularization is more specifically a squared Frobenius norm.

This section makes the following conclusions based on the above five assumptions. The linear intersection property (Assumption 2) ensures that a unique Markov boundary and a central DRS exist by Theorems 2.3.4 and 3.3.2, respectively. Theorem 3.5.2 then states that we can solve the optimization problem in eq. (5) and find a $\beta^\star$ such that $S(\beta^\star) \subseteq S_{Y|\boldsymbol{X}}$, since $\sum_{\boldsymbol{X}}$ is positive definite (3), and $E(\boldsymbol{X}|\eta^T\boldsymbol{X})$ is a linear function of $\eta^T\boldsymbol{X}$ when $Y \perp \boldsymbol{X}|\eta^T\boldsymbol{X}$ (4). Next, we can recover a nonempty subset of the Markov boundary from the non-zero coefficients of $\beta^\star$ by Theorem 3.4.2, if we also assume that $\beta^\star$ is a non-zero matrix (5). The global Markov condition (1) finally guarantees that the set of parents, children and spouses of $Y$ includes the Markov boundary of $Y$ by Theorem 2.3.6. As a result, if the above five assumptions hold, then $\beta^\star \in S_{Y|\boldsymbol{X}}$, and a nonempty subset of the parents, children and spouses of $Y$ is identifiable from $\beta^\star$. Now, if Assumption 4 does not hold for all $\eta$ such that $Y \perp \boldsymbol{X}|\eta^T\boldsymbol{X}$, then $S(\beta^\star) \subseteq S(\eta)$ for those $\eta$ which do also by Theorem 3.5.2. Notice that the above argument does not require causal faithfulness but instead requires a variant of the intersection property. Moreover, the generality of the optimization problem in eq. (5) implies that the result applies to RRLM-based regression, binary classification, and multi-class classification among possibly others.

## 3.2 The Markov boundary and variable selection

We start our investigation by extending the theoretical connection between the Markov blanket and variable selection by more precisely considering the Markov boundary. Note that, like the Markov blanket of $Y$, a trivial optimal predictor is $\boldsymbol{X}$. As a result, we can define a minimal and optimal predictor:

**Definition 3.2.1: Minimal and optimal predictor.** Let $\boldsymbol{V}$ be an optimal predictor of $Y$. If no proper subset of $\boldsymbol{V}$ satisfies the definition of an optimal predictor of $\boldsymbol{Y}$, then $\boldsymbol{V}$ is a minimal and optimal predictor of $Y$.

Similar to Theorem 2.4.2, the following theorem provides necessary and sufficient conditions for a Markov boundary to be a minimal and optimal predictor.

**Theorem 3.2.2.** If $\mathbb{M}$ is a performance metric that is maximized only when the conditional probability distribution $\mathbb{P}(Y|\boldsymbol{X})$ is estimated accurately, and $\mathbb{L}$ is a learning algorithm that can approximate any conditional probability distribution, then $\boldsymbol{M}$ is a Markov boundary of $Y$ if and only if it is a minimal and optimal predictor of $Y$.

**Proof.** First assume $\boldsymbol{M}$ is a Markov boundary of $Y$. Then, $\boldsymbol{M}$ is also a Markov blanket of $Y$, so $\boldsymbol{M}$ is an optimal predictor of $Y$ by Theorem 2.4.2. Moreover, by the definition of a Markov boundary, no proper subset of $\boldsymbol{M}$ is a Markov blanket of $Y$. As a result, no proper subset of $\boldsymbol{M}$ also satisfies the definition of an optimal predictor of $Y$. We thus conclude that $\boldsymbol{M}$ is a minimal and optimal predictor of $Y$.

The other direction follows similarly. Assume $\boldsymbol{G}$ is a minimal and optimal predictor of $Y$. Then, $\boldsymbol{G}$ is also an optimal predictor of $Y$, so $G$ is a Markov blanket of $Y$ by Theorem 2.4.2. Next, by the definition of a minimal and optimal predictor, no proper subset of $\boldsymbol{G}$ is also an optimal predictor of $Y$. Thus, no proper subset of $\boldsymbol{G}$ is also a Markov blanket of $Y$. We conclude that $\boldsymbol{G}$ is a Markov boundary of $Y$.

The above theorem implies that the Markov boundary of $Y$ is a solution to the variable selection problem when $\mathbb{P}(Y|\boldsymbol{X})$ needs to be accurately estimated. Unfortunately, this is not the case in many machine learning tasks. For example, in regression, the negative mean squared error is maximized when $E(Y|\boldsymbol{X})$ is accurately approximated – not the entire conditional distribution. Moreover, estimating the conditional expectation may require a subset of the variables in the Markov boundary of $Y$. As an example, let $\boldsymbol{X} = (X_1, X_2)$ and $Y = f(X_1) + N\left(0, \rho(X_2)^2\right)$, where $f(\cdot)$ and $\rho(\cdot)$ are some fixed functions. Then, $\mathbb{P}(Y|\boldsymbol{X})$ depends on $X_1$ and $X_2$ but $E(Y|\boldsymbol{X})$ only depends on $X_1$. This example and Theorem 3.2.2 suggest that not all RRLM methods can identify the entire Markov boundary. We will see that this intuition is indeed true in Section 3.5.

We now restrict ourselves to tasks where $\mathbb{P}(Y|\boldsymbol{X})$ is estimated and define the following:

**Definition 3.2.3: Best predictor.** A best predictor is a minimal and optimal predictor with the smallest cardinality.

Necessary and sufficient conditions regarding the Markov boundary and the best predictor can be similarly established by following the logic of the proof of Theorem 3.2.2:

**Theorem 3.2.4.** If $\mathbb{M}$ is a performance metric that is maximized only when $\mathbb{P}(Y|\boldsymbol{X})$ is estimated accurately, and $\mathbb{L}$ is a learning algorithm that can approximate any conditional probability distribution, then $\boldsymbol{M}$ is a Markov boundary of $Y$ with smallest cardinality if and only if it is a best predictor of $Y$. If only one Markov boundary exists, such as when the intersection property holds, then the Markov boundary is the best predictor and vice versa. If multiple Markov boundaries exist, then the Markov boundary or boundaries with the smallest cardinality is/are the best predictor(s) and vice versa. Thus, the best predictor can be regarded as a solution to the variable selection problem when $\mathbb{P}(Y|\boldsymbol{X})$ is estimated accurately.

## 3.3 The intersection property and SDR

We have established that the Markov boundary is an optimal variable set for a learning algorithm under certain conditions. We now link Markov boundary with SDR theory. Note that Theorem 2.3.4 provides an important sufficient condition regarding the uniqueness of a Markov boundary. At the same time, Theorem 2.5.3 provides a different sufficient condition regarding the existence of a central DRS. We connect the two ideas by first proposing the following new definition, where $\boldsymbol{Z} = (Y; \boldsymbol{X})$:

**Definition 3.3.1. (Linear intersection property).** The distribution $\mathbb{P}$ over variables $\boldsymbol{Z}$ satisfies the linear intersection property if for every $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in \mathbb{R}^{(p+1)\times d}$ where $d \leq (p+1)$ such that $\alpha_1^T\boldsymbol{Z} \perp \alpha_2^T\boldsymbol{Z} \big| \left(\alpha_3^T\boldsymbol{Z}, \alpha_4^T\boldsymbol{Z}\right)$ and $\alpha_1^T\boldsymbol{Z} \perp \alpha_4^T\boldsymbol{Z} \big| \left(\alpha_3^T\boldsymbol{Z}, \alpha_2^T\boldsymbol{Z}\right)$, we also have $\alpha_1^T\boldsymbol{Z} \perp \big| \left(\alpha_2^T\boldsymbol{Z}, \alpha_4^T\boldsymbol{Z}\right) \big| \alpha_3^T\boldsymbol{Z}$. In other words, $\alpha_1^T\boldsymbol{Z} \perp \alpha_2^T\boldsymbol{Z} \big| \left(\alpha_3^T\boldsymbol{Z}, \alpha_4^T\boldsymbol{Z}\right)$ and $\alpha_1^T\boldsymbol{Z} \perp \alpha_4^T\boldsymbol{Z} \big| \left(\alpha_3^T\boldsymbol{Z}, \alpha_2^T\boldsymbol{Z}\right) \Rightarrow \alpha_1^T\boldsymbol{Z} \perp \left(\alpha_2^T\boldsymbol{Z}, \alpha_4^T\boldsymbol{Z}\right) \big| \alpha_3^T\boldsymbol{Z}$.

Note that the intersection property is an instance of the linear intersection property, when each $\alpha_i$ is a matrix whose diagonal elements are set to zero or one, and the off-diagonal elements are all zeros. Observe that for any $\alpha \in \mathbb{R}^{(p+1)\times d}$, if $\mathbb{P}(\boldsymbol{Z})>0$, then $\mathbb{P}(\alpha^T\boldsymbol{Z})>0$. Thus, if the distribution of $\boldsymbol{Z}$ is strictly positive, then it satisfies the linear intersection property just like the original intersection property.

The following theorem guarantees the existence of a central DRS, if the linear intersection property holds:

**Theorem 3.3.2.** Let $S(\alpha)$ and $S(\phi)$ be DRSs for $Y|\boldsymbol{X}$. If the joint distribution of $(Y; \boldsymbol{X})$ satisfies the linear inters$\alpha_1=0$ection property, then $S(\alpha)\cap S(\phi)$ is also a DRS.

**Proof.** Note that a DRS must always exist, since $S(I_p)$ is a DRS. Let $S(\alpha)$ and $S(\phi)$ both be DRSs for $Y|\boldsymbol{X}$. Let $\delta$ be a basis for $S(\alpha)$ and $S(\phi)$ so that $\alpha=(\alpha_1,\delta)$ and $\phi=(\phi_1,\delta)$. Now if and $\phi_1=0$, then the conclusion is trivially true. Now let $\alpha_1\neq 0$ and $\phi_1=0$. For notational convenience, define the following:

$$\boldsymbol{W}=\begin{pmatrix}\boldsymbol{W}_1\\ \boldsymbol{W}_2\\ \boldsymbol{W}_3\end{pmatrix}=\begin{pmatrix}\alpha_1^T\boldsymbol{X}\\ \phi_1^T\boldsymbol{X}\\ \delta^T\boldsymbol{X}\end{pmatrix}.$$

Since $S(\alpha)$ and $S(\phi)$ are both DRSs, we write the following relationship between the CDFs:

$$F_{Y|\boldsymbol{W}}=F_{Y|\boldsymbol{W}_1,\boldsymbol{W}_3}=F_{Y|\boldsymbol{W}_2,\boldsymbol{W}_3}.$$

We need to also show that $F_{Y|\boldsymbol{W}}=F_{Y|\boldsymbol{W}_3}$. Note that the above relationships between the CDFs imply that $Y\perp\boldsymbol{W}|(\boldsymbol{W}_1\cup\boldsymbol{W}_3)$ and $Y\perp\boldsymbol{W}|(\boldsymbol{W}_2\cup\boldsymbol{W}_3)$ which in turn imply that $Y\perp\boldsymbol{W}_2|(\boldsymbol{W}_1\cup\boldsymbol{W}_3)$ and $Y\perp\boldsymbol{W}_1|(\boldsymbol{W}_2\cup\boldsymbol{W}_3)$. If the linear intersection property holds, we have $Y\perp\boldsymbol{W}_2|(\boldsymbol{W}_1\cup\boldsymbol{W}_3)$ and $Y\perp\boldsymbol{W}_1|(\boldsymbol{W}_2\cup\boldsymbol{W}_3)$ both imply that $Y\perp(\boldsymbol{W}_2\cup\boldsymbol{W}_1)|\boldsymbol{W}_3$. This in turn implies that $Y\perp\boldsymbol{W}|\boldsymbol{W}_3$ by Proposition 4.6 of [21]. Thus, $F_{Y|\boldsymbol{W}}=F_{Y|\boldsymbol{W}_3}$.

## 3.4 Markov boundary discovery and SDR

Recall that a strictly positive distribution is reasonable to assume whenever there is uncertainty about the data. Thus, we assume that the joint distributions under consideration satisfy the linear intersection property from here on. This implies that $\boldsymbol{M}$ is unique by Theorem 2.3.4. Without loss of generality (w.l.o.g.), we consider arranging $\boldsymbol{X}$ such that $\boldsymbol{X}=(\boldsymbol{M};\boldsymbol{X}\backslash\boldsymbol{M})$. Let $p_M$ denote the number of variables in $\boldsymbol{M}$. We can similarly consider arranging the rows of a matrix $y\in\mathbb{R}^{p\times d}$ such that $y=\left(y_{\boldsymbol{M}};y_{\boldsymbol{X}\backslash\boldsymbol{M}}\right)$, where $y_{\boldsymbol{M}}$ has $p_{\boldsymbol{M}}$ rows, and $y_{\boldsymbol{X}\backslash\boldsymbol{M}}$ has $p-p_{\boldsymbol{M}}$ rows. We claim the following:

**Theorem 3.4.1.** Suppose the joint probability distribution of $(Y;\boldsymbol{X})$ satisfies the linear intersection property. Then, there exists a central DRS $S(y_{\boldsymbol{M}})$ for $Y|\boldsymbol{M}$, and $S(y)$ is the central DRS for $Y|\boldsymbol{X}$, where $y_{\boldsymbol{X}\backslash\boldsymbol{M}}$ is a matrix of all zeros.

**Proof.** Since the joint probability distribution of $(Y;\boldsymbol{X})$ satisfies the linear intersection property, then the distribution of $(Y;\boldsymbol{M})$ also satisfies the linear intersection property (as the property holds for any four linear combinations of variables), so we know that there exists a central DRS $S(y_M)$ for $Y|\boldsymbol{M}$ such that $Y\perp\boldsymbol{M}|y_{\boldsymbol{M}}^T\boldsymbol{M}$ from Theorem 3.3.2. Together with the fact that $Y\perp\boldsymbol{X}|\boldsymbol{M}$, we conclude $Y\perp\boldsymbol{X}|y_{\boldsymbol{M}}^T\boldsymbol{M}$ from Proposition 4.6 of [21] which implies that $Y\perp\boldsymbol{X}|y^T\boldsymbol{X}$, where $y_{\boldsymbol{X}\backslash\boldsymbol{M}}$ is a matrix of zeros. Thus, $S(y)$ is a DRS for $Y|\boldsymbol{X}$. Now suppose there exists a matrix $\beta\in\mathbb{R}^{p\times d}$ such that $S(\beta)$ is also a DRS for $Y|\boldsymbol{X}$ and $S(\beta)\subset S(y)$. Then, $Y\perp\boldsymbol{X}|\beta^T\boldsymbol{X}\Rightarrow Y\perp\boldsymbol{M}|\beta^T\boldsymbol{X}$. Note that w.l.o.g. we can arrange $\beta$ so that $\beta=\left(\beta_{\boldsymbol{M}};\beta_{\boldsymbol{X}\backslash\boldsymbol{M}}\right)$. If $S(\beta)\subset S(y)$, then $\beta_{\boldsymbol{X}\backslash\boldsymbol{M}}$ must be a matrix of all zeros. We then have the contradiction $Y\perp\boldsymbol{M}|\beta_{\boldsymbol{M}}^T\boldsymbol{M}$. Thus, we must have $S(y)\subseteq S(\beta)$ for any DRS $S(\beta)$, so $S(y)$ must be the central DRS for the conditional distribution of $Y|\boldsymbol{X}$, which is guaranteed to exist by Theorem 3.3.2.

Now that we know $S(y)$ is the central DRS of $Y|\boldsymbol{X}$, we next address the question of whether we can identify a unique $\boldsymbol{M}$ from $y$. We claim that the answer is yes:

**Theorem 3.4.2.** Suppose the joint probability distribution of $(Y; \boldsymbol{X})$ satisfies the linear intersection property, and let $S(y)$ be the central DRS for $Y|\boldsymbol{X}$ as defined in Theorem 3.4.1. Let $d'$ denote the number of column dimensions in $y$. Then, we have $\sum_{i=1}^{d'} \left| y_{j,i} \right| > 0$, when the row $j$ corresponds to a variable within $\boldsymbol{M}$, and $\sum_{i=1}^{d'} \left| y_{j,i} \right| = 0$, when the row $j$ corresponds to a variable within $\boldsymbol{X} \backslash \boldsymbol{M}$.

**Proof.** The linear intersection property holds, so a central DRS $S(y)$ must always exist by Theorem 3.3.2. Also note that $S(y) = S(y_{\boldsymbol{M}})$ by Theorem 3.4.1. This implies that $\sum_{i=1}^{d'} \left| y_{j,i} \right| = 0$, when the row $j$ corresponds to a variable within $\boldsymbol{X} \backslash \boldsymbol{M}$. Next, suppose there exists a row $j$ such that $\sum_{i=1}^{d'} \left| y_{j,i} \right| = 0$, when the row corresponds to a variable within $\boldsymbol{M}$. We then have the contradiction $Y \perp \boldsymbol{X} | y_{\boldsymbol{M}}^T \boldsymbol{M} \Rightarrow Y \perp \boldsymbol{X} \left| \left( \boldsymbol{M} \backslash M_j \right) \right.$, where $M_j \in \boldsymbol{M}$.

We thus find that variables in the Markov boundary are identified by discovering the central DRS and identifying any deviations from zero in the coefficients of $y$.

## 3.5 Markov boundary discovery with ridge-regularized linear models

We now consider minimizing the following losses:

$$L_1(\alpha, \beta) = E\left\{ u\left( \alpha + \beta^T \boldsymbol{X}, Y \right) \right\}, \tag{3}$$

$$L_2(\alpha, \beta) = E\left\{ u\left( \alpha + \beta^T \boldsymbol{X}, Y \right) \right\} + \lambda tr\left( \beta^T \Sigma_X \beta \right), \tag{4}$$

where $u\,(\cdot, \cdot)$ is a convex functional, and $tr\,(\cdot)$ denotes the trace. Notice that the second loss has a covariance matrix $\Sigma_X$ added into the ridge regularization. The covariance matrix converts the ridge regularization into a new desired form whose purpose will become clear in the proof of Theorem 3.5.2.

Let $K$ denote the number the column dimensions of $\alpha$ and $\beta$, and let $k \in [1, 2, \ldots, k] = \mathbb{K}$. We first provide the following analysis of eq. (3):

**Theorem 3.5.1.** Assume $\Sigma_X$ is positive definite, and let $S(\eta)$ be any DRS such that $E\left( \boldsymbol{X} | \eta^T \boldsymbol{X} \right)$ is a linear function of $\eta^T \boldsymbol{X}$. If $\left( \alpha^\star, \beta^\star \right)$ minimizes eq. (3) and $\beta^\star$ is unique, then $S\left( \beta_k^\star \right) \subseteq S(\eta)$ for all $k \in \mathbb{K}$.

**Proof.** First, when $\eta = I_p$, then $E\left( \boldsymbol{X} | \eta^T \boldsymbol{X} \right)$ is always of linear function of $\boldsymbol{X}$. Also, we have:

$$\begin{aligned} E\left[ u\left( \alpha + \beta^T \boldsymbol{X}, Y \right) \right] &= E_{Y, \eta^T \boldsymbol{X}} E_{\boldsymbol{X} | Y, \eta^T \boldsymbol{X}} \left[ u\left( \alpha + \beta^T \boldsymbol{X}, Y \right) \right] \\ &= E_{Y, \eta^T \boldsymbol{X}} E_{\boldsymbol{X} | \eta^T \boldsymbol{X}} \left[ u\left( \alpha + \beta^T \boldsymbol{X}, Y \right) \right], \end{aligned}$$

where the last equality follows since $Y \perp \boldsymbol{X} | \eta^T \boldsymbol{X}$. Recall that $u(\ \cdot,\ \cdot)$ is convex, so we apply Jensen's inequality as follows:

$$E_{Y, \eta^T \boldsymbol{X}} E_{\boldsymbol{X} | \eta^T \boldsymbol{X}} \left[ u\left( \alpha + \beta^T \boldsymbol{X}, Y \right) \right] \geq E_{Y, \eta^T \boldsymbol{X}} \left[ u\left( \alpha + \beta^T E\left( \eta^T \boldsymbol{X} \right), Y \right) \right]$$

W.l.o.g., we assume that $E(\boldsymbol{X}) = 0$. Additionally, since $E\left( \eta^T \boldsymbol{X} \right)$ is a linear function of $\eta^T \boldsymbol{X}$, we apply Proposition 4.2 of [21] so that:

$$L_1(\alpha, \beta) \geq L_1\left( \alpha, P_\eta\left( \Sigma_{\boldsymbol{X}} \right) \beta \right),$$

where $P_\eta\left( \Sigma_{\boldsymbol{X}} \right)$ is the projection matrix $\eta\left( \eta^T \Sigma_{\boldsymbol{X}} \eta \right)^{-1} \eta^T \Sigma_{\boldsymbol{X}}$. The conclusion now follows because the solution $\beta^\star$ is unique. ■

A similar argument is given in [24] and [21], but when $E(\boldsymbol{X}|\eta^T\boldsymbol{X})$ is a linear function of all $\eta^T\boldsymbol{X}$ such that $Y \perp \boldsymbol{X}|\eta^T\boldsymbol{X}$, and when $K=1$.

Note that the assumption that $E(\boldsymbol{X}|\eta^T\boldsymbol{X})$ is a linear function of $\eta^T\boldsymbol{X}$ may not always hold. However, it usually holds when $p$ is large [13] and always holds when the distribution of $\boldsymbol{X}$ is elliptically symmetric [25]. The above theorem thus implies an important point: we can find a subset of the minimum DRS $S(\eta)$ such that $E(X|\eta^T\boldsymbol{X})$ is still a linear function using eq. (3), provided that $\beta^\star$ is unique. Moreover, $\beta^\star$ is guaranteed to include a subset of the Markov boundary if $E(\boldsymbol{X}|\eta^T\boldsymbol{X})$ is a linear function when $S(\eta)=S_{Y|\boldsymbol{X}}$ in the sense of Theorem 3.4.2.

We now claim that we can drop the assumption that $\beta^\star$ is unique by adding ridge regularization with a covariance matrix as in eq. (4):

**Theorem 3.5.2.** Assume $\Sigma_{\boldsymbol{X}}$ is positive definite, and let $S(\eta)$ be any DRS such $E(\boldsymbol{X}|\eta^T\boldsymbol{X})$ that is a linear function of $\eta^T\boldsymbol{X}$. If $(\alpha^\star, \beta^\star)$ minimizes eq. (4), then $S\left(\beta_k^\star\right) \subseteq S(\eta)$ for all $k \in \mathbb{K}$.

**Proof.** The first part of the proof is similar to Theorem 3.5.1. We have:

$$E\left[u\left(\alpha+\beta^T\boldsymbol{X}, Y\right)\right] = E_{Y,\eta^T\boldsymbol{X}}E_{\boldsymbol{X}|\eta^T\boldsymbol{X}}\left[u\left(\alpha+\beta^T\boldsymbol{X}, Y\right)\right].$$

Applying Jensen's inequality:

$$E_{Y,\eta T_X}E_{\boldsymbol{X}|\eta T_X}\left[u\left(\alpha+\beta^T\boldsymbol{X}, Y\right)\right] \geq E_{Y,\eta^T\boldsymbol{X}}\left[u\left(\alpha+\beta^T\left(X|\eta^T\boldsymbol{X}\right), Y\right)\right]. \tag{5}$$

Also consider, for any $k \in [1, 2, \ldots K]$:

$$\begin{aligned} var\left(\beta_k^T\boldsymbol{X}\right) &= var\left(E\left(\beta_k^T\boldsymbol{X}|\eta^T\boldsymbol{X}\right)\right) + E\left(var\left(\beta_k^T\boldsymbol{X}|\eta^T\boldsymbol{X}\right)\right) \\ &\geq var\left(E\left(\beta_k^T\boldsymbol{X}|\eta^T\boldsymbol{X}\right)\right). \end{aligned} \tag{6}$$

Putting eqs (5) and (6) together, we have:

$$L_2(\alpha,\beta) \geq E_{Y,\eta^T\boldsymbol{X}}\left[u\left(\alpha+\beta^T E(\boldsymbol{X}|\eta^T\boldsymbol{X}), Y\right)\right] + \lambda\sum_{k=1}^{K} var\left(\beta_k^T E(\boldsymbol{X}|\eta^T\boldsymbol{X})\right). \tag{7}$$

W.l.o.g., we assume that $E(\boldsymbol{X})=0$. Additionally, since $E(\boldsymbol{X}|\eta^T\boldsymbol{X})$ is a linear function of $\eta^T\boldsymbol{X}$, we apply Proposition 4.2 of [21] so that:

$$L_2(\alpha,\beta) \geq L_2(\alpha, P_\eta(\Sigma_{\boldsymbol{X}})\beta),$$

where $P_\eta(\Sigma_{\boldsymbol{X}}) = \eta\left(\eta^T\Sigma_{\boldsymbol{X}}\eta\right)^{-1}\eta^T\Sigma_{\boldsymbol{X}}$.

Now let $S(\eta)$ be the minimum DRS such that $E(\boldsymbol{X}|\eta^T\boldsymbol{X})$ is still a linear function of $\eta^T\boldsymbol{X}$. If $S(\beta_k) \supset S(\eta)$, then $var\left(E\left(\beta_k^T\boldsymbol{X}|\eta^T\boldsymbol{X}\right)\right) > 0$. This implies that eqs (6) and (7) are strict inequalities. As a result, such a $\beta_k$ cannot be used to minimize eq. (4). ■

The above proof is a modification of Theorem 1 in [26] which was used in the context of binary SVMs. We believe the above theorem is powerful, since it applies to all convex losses regardless of the nonlinearities in the functional relationship.

The conclusions summarized in Section 3.1 follow from a straightforward synthesis of the above theorems. The linear intersection property ensures that a unique Markov boundary and a central DRS exist by Theorems 2.3.4 and 3.3.2, respectively. Since $\Sigma_{\boldsymbol{X}}$ is positive definite, and $E(\boldsymbol{X}|\eta^T\boldsymbol{X})$ is a linear function of $\eta^T\boldsymbol{X}$ when $Y \perp \boldsymbol{X}|\eta^T\boldsymbol{X}$, we can solve eq. (4) and find a $\beta^\star$ such that $S\left(\beta_k^\star\right) \subseteq S_{Y|\boldsymbol{X}}$ for all $k \in \mathbb{K}$ by Theorem 3.5.2. We finally recover a subset of the parents, children and spouses of $Y$ from the non-zero coefficients of $\beta^\star$ by Theorems 3.4.2 and 2.3.6, assuming $\beta^\star$ is a non-zero matrix which is almost always the case in practice.

# 4 Experiments

## 4.1 Implementation

We now describe an implementation of the above ideas using the empirical version of eq. (4). Let $Y_n$ and $X_n$ denote two fixed $p$ by $n$ matrices of $n$ samples from the joint distribution of $(Y; \boldsymbol{X})$. The empirical version of eq. (4) can be written as follows:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{u\left(\alpha+\beta^T X_{n,i}, Y_{n,i}\right)\right\}+\lambda tr\left(\beta^T \hat{\Sigma}_X \beta\right). \tag{8}$$

If the empirical covariance matrix $\hat{\Sigma}_{\boldsymbol{X}}$ is non-singular, then eq. (8) can be minimized with standard packages for RRLMs such as Glmnet [11] or LIBSVM [12] by standardizing $X_n$ so that $Z_n = \hat{\Sigma}_{\boldsymbol{X}}^{-1/2}\left(X_n - \bar{X}_n\right)$. Then eq. (8) becomes:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{u\left(\alpha+y^T Z_{n,i}, Y_{n,i}\right)\right\}+\lambda tr\left(y^T y\right), \tag{9}$$

where $y = \hat{\Sigma}_{\boldsymbol{X}}^{-1/2}\beta$. We then recover $\hat{\beta}^{\star}$ by $\hat{\beta}^{\star} = \hat{\Sigma}_{\boldsymbol{X}}^{-1/2}\hat{y}^{\star}$.

We propose to solve eq. (9) and use permutation analysis to recover the p-values of each of the coefficients in $\hat{\beta}^{\star}$. When the number of distinct samples is fewer than the number of dimensions $p$, we use the method of shrunken covariance estimators to estimate the covariance matrix [27, 28]. From here on, we use the term Covariance Ridge with Permutation (CRP) to refer to a permutation analysis by repetitively solving eq. (9) in order to recover coefficient specific p-values.

## 4.2 Toy examples

We provide two toy simulations demonstrating the new theory. First, we obtained 1,000 samples 100 times from 5 independent random variables with normal distributions with standard deviations that were drawn from a standard normal. We then created the random variable $Y$ using the following fifth order polynomial:

$$Y = \varepsilon + \theta_0 + \sum_{i=1}^{5}\theta_i X_i^i,$$

where the coefficients $\theta$ were also drawn from a standard normal, and $\varepsilon$ is a normally distributed error with small standard deviation 1E-4 such that $\varepsilon \perp (X_1, \ldots, X_5)$. We then added l, 5, 15, 45, and 95 additional independent variables with normal distributions also with standard deviations drawn from a standard normal. The five variables $(X_1, \ldots, X_5)$ thus comprise the Markov boundary of $Y$, and the additional variables are not part of the Markov boundary of $Y$.

CRP identified a subset of the five Markov boundary variables in all one hundred replicates across all numbers of additional variables. The mean number of correctly identified Markov boundary variables was 2.11 (SD: 0.751), and there was no significant decrease in the number of correctly identified Markov boundary variables as the number of additional variables increased. This is expected because the multivariate normal is elliptically symmetric, so $E\left(\boldsymbol{X}|\eta^T\boldsymbol{X}\right)$ is a linear function of any $\eta^T\boldsymbol{X}$. Hence, the optimal $\beta^{\star}$ obtained from minimizing eq. (4) is guaranteed to contain a subset of the Markov boundary from Theorem 3.5.2. For additional empirical evidence, we also ran a similar experiment, where the variables were drawn from a beta (2,2) distribution; recall that this distribution is also elliptically symmetric. We obtained similar results, except with a higher mean number of correctly identified variables at 4.47 (0.028).

In the next experiment, the explanatory variables were drawn from beta distributions, where the alpha and beta parameters were chosen by sampling from a uniform distribution between 0 and 5. Most instantiations of the beta distribution are not elliptically symmetric, but $E(\boldsymbol{X}|\eta^T\boldsymbol{X})$ is almost always a linear function in the high dimensional setting, so we expect eq. (4) to still perform reasonably well even when we increase the number of additional variables. Nonetheless, we also expect that the method will not perform as well as with the beta(2,2) distributed variables. Indeed, we found that the mean number of correctly identified variables dropped to 4.18 (0.061) but again with no significant degradation as the number of additional variables increased.

## 4.3 Expert-designed models

We evaluated CRP on datasets generated from four expert-designed discrete Bayesian networks (brief descriptions are given in Table 1). Note that many of the relationships between the variables in these networks are nonlinear. We compared the performances of the following three methods:

(1) The HITON-PC algorithm is a parent and child discovery method [29]. Even though this method is not a Markov boundary discovery algorithm per se, it outperforms many existing algorithms on several metrics in identifying the true Markov boundary [30, 31]. We used the Causal Explorer implementation [32] with the $G^2$ test. Additionally, we selected the $\alpha$ hyperparameter from the set {0.001, 0.01, 0.05, 0.10} using 5-fold cross-validation with an RBF-kernel SVM. The SVM's $C$ and $\sigma$ hyperparameters were in turn chosen within the folds from {0.1, 1, 10 100} and the median distance between samples multiplied by {2/3, 0.8, 1, 1.25, 1.5}, respectively. Finally, we set the $k$ hyperparameter for HITON-PC to 2, the largest value we could use so that the algorithm completed in a reasonable length of time in our experiments. In practice, values of $k$ up to 4 are recommended for HITON-PC [30].

(2) The Kernel Backward Elimination (KBE) algorithm is, to our knowledge, the best performing Markov boundary *ranking* method [33]. Note that ranking may be more useful than selection when the number of selected variables is large. We used the default hyperparameter settings of $\sigma$ set to the median distance between samples and the ridge regularization set to 1E-4.

(3) The CRP algorithm was implemented by solving eq. (9) using the Glmnet package [11]. We used a multinomial logistic loss or the MSE loss, when the former failed due to insufficient samples per group. Hyperparameters were determined by the cvglmnet function as provided in the package, and 1,000 permutations were used to obtain approximate p-values.

**Table 1:** Descriptions of expert-designed discrete Bayesian networks. The fourth column gives the minimum and maximum cardinality of the parent and child set for each network.

| Network | Num. of Variables | Num. of Edges | Min/Max \|PC\| |
|---|---:|---:|---:|
| Child10 | 200 | 257 | l/8 |
| Alarm10 | 270 | 570 | l/9 |
| Pigs | 441 | 592 | l/41 |
| Gene | 801 | 972 | 0/11 |

In this experiment, our goal was to assess the ability of the above three algorithms in identifying a subset of the Markov boundary. Note that HITON-PC and KBE are guaranteed to discover the Markov boundary of any response in the infinite sample limit under their respective assumptions. However, CRP can discover only a subset of the Markov boundary of some unknown cardinality as noted in Theorem 3.5.2. We thus chose to evaluate the algorithms in their ability to discover a single variable in the Markov boundary. To do this, we randomly chose 50 variables as the responses from each Bayesian network, ran the three algorithms using sample sizes of {50, 100, 200, 300, 400, 500}, and then counted the number of times out of 50 that each algorithm correctly identified any Markov boundary variable. We repeated the experiments for each of the 4 networks and thus ran a total of $50 \times 4 = 200$ experiments per sample size. Recall that KBE and CRP output an ordering of the variables (the latter in terms of the smallest to largest p-values), so we only evaluated the $h$ lowest ranked variables for these two methods, where $h$ is the cardinality of the output from HITON-PC. We did not simply impose a $p \le 0.05$ cut-off for CRP for reasons that will become clear in Section 4.5. The results of the experiments across multiple sample sizes are summarized in Figure 1.

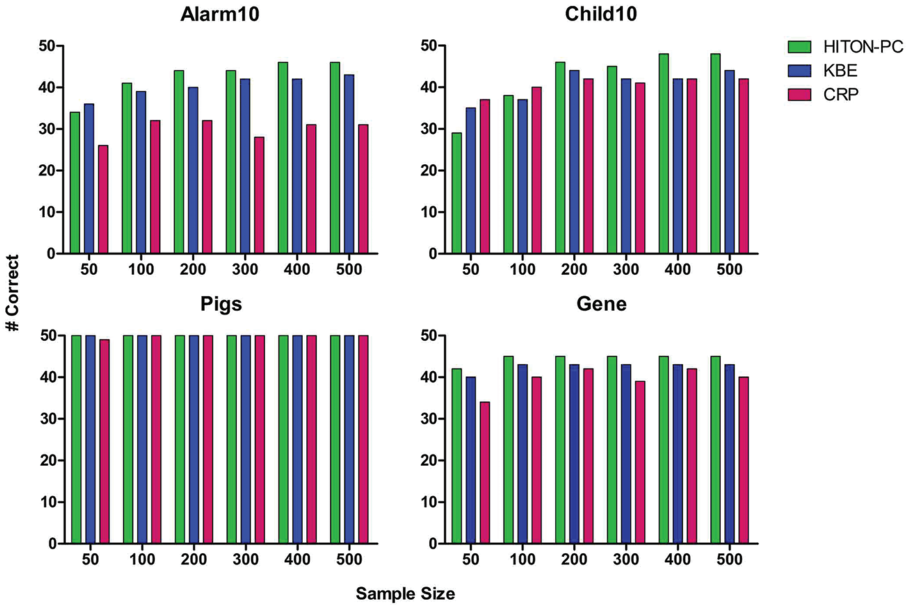

**Figure 1:** The number of times HITON-PC, KBE, and CRP correctly detected a single variable in 50 Markov boundaries across four expert-designed discrete Bayesian networks. HITON-PC and KBE outperform CRP by a small margin across most sample sizes.

To evaluate differences in performance regardless of sample size, we counted the number of times a Markov boundary variable was identified across all sample sizes and performed two-tailed Fisher's exact test. We used sample sizes of {50, 100, 200, 300, 400, 500} because these sample sizes are typical of the HDLSS datasets in biology to which RRLMs are applied. We found that HITON-PC and KBE outperformed CRP (both $p < 0.001$) using the Bonferroni-adjusted threshold of $0.05/2 = 0.025$. This result held even when we controlled for sample size and network using a $G^2$ test ($G^2 = 81.67$, $p < 0.001$ with HITON-PC; $G^2 = 39.51$, $p = 0.004$ with KBE). The results therefore suggest that CRP underperforms the nonparametric methods. However, examining the histograms in Figure 1 reveals that CRP is usually only outperformed by a small margin in the majority of sample sizes despite "only being linear."

## 4.4 Linear Gaussian non-recursive structural equation models

Some real-world causal networks are likely to be cyclic or contain continuous variables. Examples include networks based on gene expression, where local causal discovery methods are most often applied. As a result, we evaluated CRP on datasets generated from linear Gaussian non-recursive structural equation models with independent errors (SEM-IEs), since our theoretical conclusions as summarized in Section 3.1 hold regardless of cyclicity. Moreover, it is known that linear Gaussian SEM-IEs and their stationary distributions (if they exist) satisfy the global Markov condition [34]. Recall also that the multivariate Gaussian distribution is strictly positive and therefore satisfies the intersection property by Proposition 2.2.2. The Markov boundary of the response in a stationary linear Gaussian SEM-IE thus uniquely includes the parents, children and spouses because these variables (but no proper subset of them) d-separate all other variables from the response.

We generated four networks using the default settings and default randomization methods in the TETRAD software package [35]. Moreover, unlike in Section 4.3, we equipped HITON-PC with the z-test (using Fisher's r-to-z transformation with the partial autocorrelation coefficient), and KBE with linear reproducing kernels due to the linearities in the network. Hyperparameters for HITON-PC were chosen by 5-fold cross validation using ridge regression with the best $\lambda$ obtained from the set {1E-6, 1E-4, 1E-2} as assessed within the folds. We also used the MSE loss for CRP. Both the task and the additional settings were otherwise kept the same as in Section 4.3. We therefore performed a total of 200 experiments per sample size. Note that the multivariate Gaussian distribution is elliptically symmetric, so we expect CRP to perform well, since its solution space is guaranteed to only contain a subset of the Markov boundary in the large sample limit by Theorem 3.5.2. Results are summarized in Figure 2 and indeed show that CRP's performance is competitive. CRP even outperforms HITON-PC (but not KBE) by a small margin by two-tailed Fisher's exact test ($p = 0.008$ with HITON-PC; $p = 0.205$ with KBE) and by $G^2$ test controlling for sample size and network ($G^2 = 40.49$, $p = 0.014$ with HITON-PC; $G^2 = 12.07$, $p = 0.883$ with KBE).

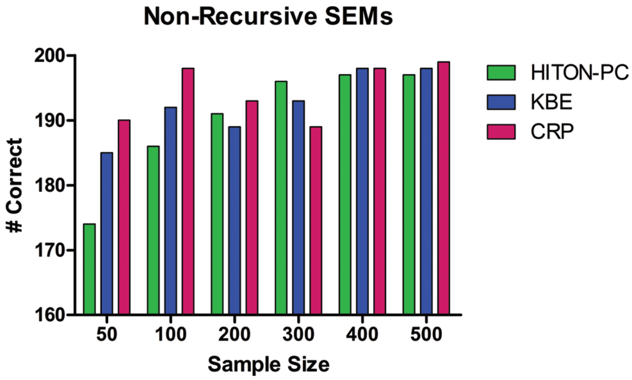

**Figure 2:** Same as Figure 1 except the 50 Markov boundaries were obtained from four linear Gaussian non-recursive SEM-IEs. The results are combined across the four SEM-IEs. CRP outperforms HITON-PC by a small margin in detecting a single variable within the Markov boundary.

## 4.5 Application to gene expression data

In the final set of experiments, we applied the algorithms to two real gene-expression datasets titled NOTCH1 and RELA [36]. These datasets are valuable in that they come with gold standard solution sets consisting of experimentally verified genes that are downstream of the NOTCH1 and RELA transcription factors. The downstream genes were specifically identified using the following procedure. First, the samples were randomized to either a control or experimental group, where the transcription factor of interest was knocked down (e.g., by siRNA). A t-test was then performed with alpha level set to 0.05 to identify differentially expressed genes between the two groups. Second, the set of genes with promoter region-transcription factor binding was identified using genome-wide binding data (ChiP-on-chip for NOTCH1, and ChIP-sequencing for RELA). The final silver standard set of genes was obtained by overlapping the set of genes from the experimental knockdown data and the set of genes from the binding data.

Our goal was to compare the algorithms in their ability to identify the downstream genes of the transcription factors. We were also interested in the ease to which the algorithms' outputs could be interpreted by humans in order to help prioritize experimentation. We ran HITON-PC on the two datasets using the $G^2$ test (as suggested in [30]), where each variable was first discretized into three bins. The $\alpha$ hyperparameter was chosen using 5-fold cross validation from {0.001, 0.01, 0.05, 0.10} using RBF kernel ridge regression with $\lambda$ from {1E-6, 1E-4, 1E-2} and $\sigma$ from the median distance between samples multiplied by {2/3, 0.8, 1, 1.25, 1.5}. HITON-PC ultimately provided an output of 1,530 variables for NOTCHl and 15,375 variables for RELA. Similarly, CRP using an MSE loss and a $p \leq 0.05$ cut-off provided an output of 1,045 variables for NOTCH1 and 921 variables for RELA. Interpreting such large sets of genes is clearly impractical for humans, and experimentally verifying them is even more impractical. Of course a simple heuristic solution would be to decrease the alpha threshold until the output contains a reasonable amount of variables. On the other hand, an alternative and perhaps a more principled strategy is to provide to the experimenter an ordered sequence of variables instead of an unordered set of variables such that the lowest ranked variables have the highest priority for experimental manipulation. We therefore compared KBE and CRP in detail, as both algorithms output variable rankings.

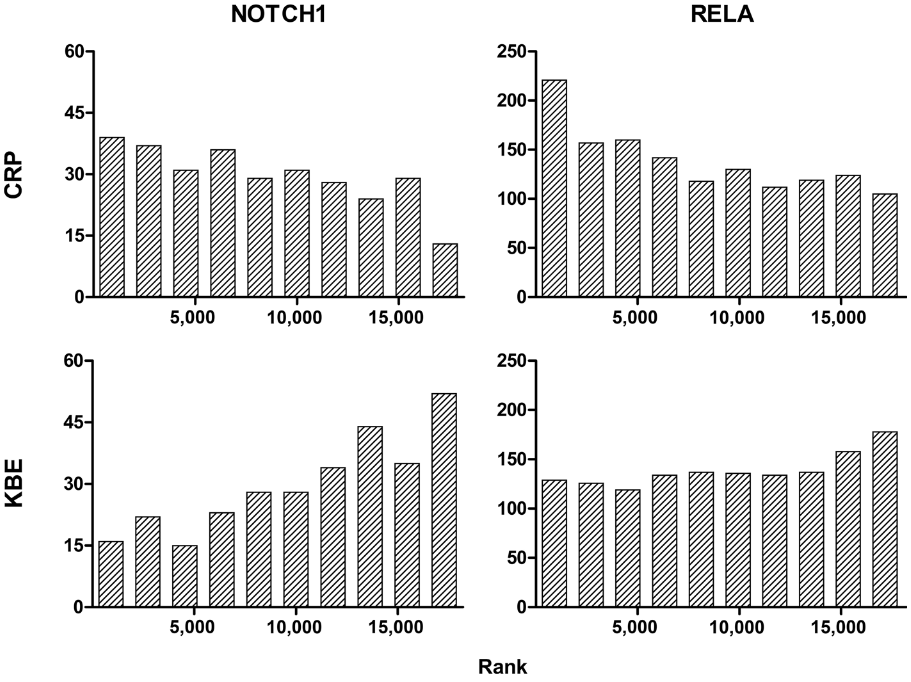

**Figure 3:** Histograms of the rankings obtained from CRP (top row) and KBE (bottom row) for the NOTCH1 (first column) and RELA (second column) datasets. CRP outperforms KBE, since the histograms for CRP are more right skewed than those for KBE.

The KBE algorithm was run as described in Section 4.3, and the CRP algorithm was run with the MSE loss. Histograms of the rankings for the silver standard genes are shown in Figure 3. Note that the ideal output corresponds to a single probability mass at rank 1, and thus a more right skewed distribution denotes better performance. The histograms show that CRP outperforms KBE by a large margin, since the distributions of the rankings for CRP are more right skewed than those for KBE. Note that the histograms created from the output of KBE using a linear reproducing kernel were also left skewed.

## 5 Discussion

We have shown that the Markov boundary is an optimal solution to the feature selection problem for learners which must approximate the conditional distribution of $Y$ given $\boldsymbol{X}$. However, empirical evidence suggests that identifying the Markov boundary in its entirety can be difficult for some problems involving high dimensional data. We can thus instead choose to identify a subset of the Markov boundary. By connecting Markov boundary and sufficient dimension reduction theory, we have shown that a modified form of RRLMs can get very close to identifying a subset of the Markov boundary. This fact holds under a variety of losses and nonlinearities in the functional relationships between the response and explanatory variables. In practice, the modified RRLMs compare favorably to state-of-the-art full Markov boundary discovery methods in both performance and interpretability on an experimentally verifiable task with two gene expression datasets.

The CRP algorithm may have outperformed KBE on gene expression data due to a different reason besides attempting to solve an easier problem. Specifically, ranking variables by their p-values may in practice be superior than ranking them by their residual errors (when removed). This preliminary claim is partially supported by the observation that CRP also outperforms KBE with linear reproducing kernels which is in fact equivalent to linear ridge regression [37, 38]. Naturally, one may next wonder whether combining p-values with backward elimination can yield even better performance. Unfortunately, obtaining p-values at every iteration with a backward elimination method is often too computationally expensive in practice, so this strategy is not tractable unless the asymptotic distributions of the coefficients can be predefined under the null.

Besides its advantages, CRP also has several limitations. First, we must emphasize that this method does not replace Markov boundary discovery methods which are guaranteed to discover the Markov boundary in its entirety in the infinite sample limit (under their respective assumptions). Nonetheless, we hope that the ideas presented in this paper suggest that RRLMs are more useful than previously demonstrated. Second, we were unable to run the method on high-dimensional datasets with several tens of thousands of variables on desktop computers due to the memory requirements of the covariance matrix. This limits the applicability of the method, and we are not currently aware of ways to alleviate the issue. Third, the application of the method to gene expression data may seem premature given current understanding about d-separation in cyclical causal models. Indeed, the global Markov condition is only known to apply to the linear Gaussian as well as to discrete SEM-IEs when cyclicity exists [34, 39]. However, the method appears to perform well in practice, either because the global Markov condition holds in the tested cases or due to other reasons. For example, Spirtes defined another separation condition which holds more generally with non-linear SEM-IEs via d-separation in *collapsed graphs*, which we now call *cd-separation* for collapsed d-separation [34]. Unfortunately, cd-separation may not be the finest separation condition across all stationary distributions, in the sense that there are separation conditions which may entail additional conditional independencies; indeed, cd-separation implies d-separation but the converse is not true, so cd-separation is coarser than d-separation. Equivalently, we can state that d-connection implies cd-connection. We believe that the modified RRLMs may, in the worst case, be discovering a subset of those variables which are cd-connected with the response. As a result, the modified RRLMs are also correctly discovering some of those variables which are d-connected with the response.

A variety of interesting open questions remain. We do not know if the same conclusions hold if the covariance matrix is dropped from the ridge, or if the lasso is used instead. Moreover, it may be interesting to know whether certain losses have stronger theoretical guarantees than others in identifying a larger portion of the Markov boundary. Nonlinear methods such as those based on reproducing kernels may further allow us to drop some of the assumptions in Theorem 3.5.2. Finally, it may be worthwhile to explore violations of the intersection property, where a central DRS may not exist. We conjecture that the solution to a modified RRLM will be located within the union of all minimal DRSs and thus can be used to identify a subset of the union of all Markov boundaries.

In conclusion, we have presented theoretical and experimental results concerning RRLMs that strengthen their interpretation as a causal discovery method. Although modified RRLMs are not guaranteed to detect a subset of the Markov boundary in the large sample limit, they almost do so in theory and often do so in practice.

**Acknowledgments:** The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Library of Medicine.

**Funding**: Research reported in this publication was supported by grant U54HG008540 awarded by the National Human Genome Research Institute through funds provided by the trans-NIH Big Data to Knowledge initiative. The research was also supported by the National Library of Medicine of the National Institutes of Health under award numbers T15LM007059 and R01LM012095.

# References

1. Neapolitan RE. Learning Bayesian networks, Prentice Hall series in artificial intelligence. Upper Saddle River, NJ: Pearson Prentice Hall, 2004.
2. Tsamardinos I, Aliferis CF. Towards principled feature selection: relevancy, filters and wrappers. Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics. 2003.
3. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference, representation and reasoning. San Mateo, CA: Morgan Kaufmann, 1988.
4. Statnikov A, Lytkin NI, Lemeire J, Aliferis CF. Algorithms for discovery of multiple Markov boundaries. J Mach Learn Res 2013;14:499–566.
5. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genomewide expression patterns. Proc Nat Acad Sci 1998;95:14863–8.
6. Holmes JH, Durbin DR, Winston FK. The learning classifier system: an evolutionary computation approach to knowledge discovery in epidemiologic surveillance. Artif Intell Med 2000;19:53–74.
7. Li L, Weinberg CR, Darden TA, Pedersen LG. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the Ga/Knn method. Bioinformatics 2001;17:1131–42.
8. Zhou X, Kao MCJ, Wong WH. Transitive functional annotation by shortest-path analysis of gene expression data. Proc Nat Acad Sci 2002;99:12783–8.
9. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 1970;12:55–67.
10. Vapnik V, Chapelle O. Bounds on error expectation for support vector machines. Neural Comput 2000;12:2013–36.
11. Qian J, Hastie T, Friedman J, Tibshirani R, Simon N. 2013. *Glmnet for Matlab* 2013. Available at http://www.stanford.edu/~hastie/glmnet matlab/
12. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2011;2.
13. Hall P, Li KC. On almost linearity of low dimensional projections from high dimensional data. Ann Stat 1993;21:867–89.
14. Dawid AP. Conditional independence in statistical theory. J Roy Stat Soc: Ser B 1979;41:1–31.
15. Peters J. On the intersection property of conditional independence and its application to causal discovery. J Causal Infer 2014;3:97–108.
16. Lemeire J, Dominik J. Replacing causal faithfulness with algorithmic independence of conditionals. Minds Mach 2010;23:227–49.
17. Lemeire J, Meganck S, Cartella F. 2010. Robust independence-based causal structure learning in absence of adjacency faithfulness. Proceedings of the Fifth European Workshop on Probabilistic Graphical Models.
18. Dougherty E, Brun M. On the number of close-to-optimal feature sets. Cancer Inf 2006;2:189–96.
19. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Nat Acad Sci USA 2006;103:5923–8.
20. Statnikov A, Aliferis CF. Analysis and computational dissection of molecular signature multiplicity. PLoS Comput Biol 2010;65:el000790.
21. Cook DR. Regression graphics: ideas for studying regressions through graphics. Edited by Wiley Series in Probability and Statistics. Canada: John Wiley & Sons, 1998.
22. Richardson TS, Spirtes P. Automated discovery of linear feedback models. In: Glymour C, and Cooper G, editors. Computation, causation, and discovery. Menlo Park: AAAI Press, 253–302.
23. Richardson TS, Spirtes P. Ancestral graph Markov models. Ann Stat 2002;30:962–1030.
24. Duan N, Li K. Slicing regression: a link-free regression method. Ann Stat 1991;19:505–30.
25. Eaton ML. A characterization of spherical distributions. J Multivariate Anal 1986;20:272–6.
26. Li B, Artemiou A, Li L. Principal support vector machines for linear and nonlinear sufficient dimension reduction. Ann Stat 2011;39:3182–210.
27. Ledoit O, Wolf M. Honey, I shrunk the sample covariance matrix. J Portfolio Manag 2004;30:110–19.
28. Ledoit O, Wolf M. Nonlinear shrinkage estimation of large-dimensional covariance matrices. Ann Stat 2012;40:l024–l060.
29. Aliferis CF, Tsamardinos I, Statnikov A. HITON: a novel Markov blanket algorithm for optimal variable selection. AMIA 2003 *Annual Symposium Proceedings*, 2003:21–5.
30. Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD. Local causal and Markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. J Mach Learn Res 2010a;11:171–234.
31. Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD. Local causal and Markov blanket induction for causal discovery and feature selection for classification part ii: analysis and extensions. J Mach Learn Res 2010b;11:235–84.
32. Statnikov A, Lytkin A, Lemeire J, Aliferis CF. Causal explorer: a Matlab library of algorithms for causal discovery and variable selection for classification. In: Guyon I, Aliferis CF, Cooper GF, Elisseeff A, Pellet JP, Spirtes P and Statnikov A, editors. Challenges in machine learning. Volume 2: causation and prediction challenge. Bookline, MA: Microtome Publishing, 2010.

33. Strobl Ev, Visweswaran S. Markov Banket Ranking using Kernel-based Conditional Dependence Measures. NIPS Workshop on Causality. 2013.
34. Spirtes P. Directed cyclic graphical representations of feedback models. Uncertainty Artif Intell 1995:491–8.
35. Scheines R, Spirtes P, Glymour C, Meek C, Richardson T. The TETRAD project: constraint based aids to causal model specification. Multivariate Behav Res 1998;33:65–117.
36. Statnikov A, Henaff M, Lytkin NI, Aliferis CF. New methods for separating causes from effects in genomics data. BMC Genomics 2012;13:S22. doi: 10.1186/1471-2164-13-S8-S22. Epub 2012 Dec 17.
37. Fukumizu K, Bach FR, Jordan MI. Kernel dimension reduction in regression. Ann Stat 2009;37:1871–905.
38. Fukumizu K, Leng C. Gradient-based Kernel dimension reduction for regression. J Am Stat Assoc 2014;109:359–70.
39. Pearl J, Dechter R. Identifying independencies in causal graphs with feedback. Uncertainty in Artificial Intelligence: Proceedings of the Twelfth Conference, 1996: 420–426.
40. Chapelle O. Training a support vector machine in the primal. Neural Comput 2007;19:1155–78.
41. Cortes C, Vapnik V. Support vector networks. Mach Learn 1995;20:273–97.
42. Wu Y, Liu Y. Robust truncated loss support vector machines. J Am Stat Assoc 2007;102:974–83.