ALGORITHMS FOR DISCOVERY OF MULTIPLE MARKOV BOUNDARIES: APPLICATION TO THE MOLECULAR SIGNATURE MULTIPLICITY PROBLEM

By

Alexander Romanovich Statnikov

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Informatics

December, 2008

Nashville, Tennessee

Approved:

Professor Constantin F. Aliferis

Professor Gregory F. Cooper

Professor Douglas P. Hardin

Professor Daniel R. Masys

Professor Ioannis Tsamardinos

To mankind and science

ACKNOWLEDGEMENTS

I am especially indebted to my academic advisor, Dr. Constantin F. Aliferis. Without his colossal commitment and contribution to this project, this work would probably never exist. I would also like to acknowledge members of my Ph.D. committee, Dr. Gregory F. Cooper, Dr. Douglas P. Hardin, Dr. Daniel R. Masys, and Dr. Ioannis Tsamardinos, who were very supportive of my research and contributed significantly to it. I am also grateful to Dr. Dean Billheimer, Laura E. Brown, Dr. Frank E. Harrell, Dr. Isabelle Guyon, and Dr. Subramani Mani for providing me with feedback, advice, references, and software relevant to my project. In addition, I would like to acknowledge Dr. Cynthia S. Gadd and Dr. Nancy Lorenzi for their ongoing support of my studies and research in the Biomedical Informatics Ph.D. program.

Finally, I would like to express my gratitude to my wife, Kristina Statnikova, my son, Grigory Statnikov, my parents, Roman Statnikov and Yelena Feofilova, my sister, Irina Statnikova, and my mother-in-law, Ludmila Morozyuk. Their encouragement and support were crucial for my Ph.D. studies and dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Preamble

The problem of variable/feature selection is of fundamental importance in machine learning and applied statistics, especially when it comes to analysis, modeling, and discovery from high-dimensional data (Guyon and Elisseeff, 2003; Kohavi and John, 1997). 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 data-generative process (Guyon and Elisseeff, 2003). An emerging class of algorithms proposes a principled solution to the variable selection problem by identification of a Markov blanket of the response variable of interest (Aliferis et al., 2008a; Aliferis et al., 2003; Tsamardinos and Aliferis, 2003; Tsamardinos et al., 2003b). A *Markov blanket* is a set of variables conditioned on which all the remaining variables excluding the response variable are statistically independent of the response variable. Under assumptions about the learner and loss function, a Markov blanket is the solution to the variable selection problem (Tsamardinos and Aliferis, 2003). A related useful concept is *Markov boundary* (or non-redundant Markov blanket) that is a Markov blanket such that no proper subset of it is a Markov blanket.

An important theoretical result states that if the distribution satisfies the intersection property, then it is guaranteed to have a unique Markov boundary of the response variable (Pearl, 1988). However, many real-life distributions contain multiple Markov boundaries and violate the intersection property. For example, the multiplicity of molecular signatures (Azuaje and Dopazo, 2005; Somorjai et al., 2003), a phenomenon ubiquitous in analysis of high-throughput molecular data, suggests existence of multiple Markov boundaries in these distributions.

There are at least two practical benefits of an algorithm that could systematically extract all Markov boundaries of the response variable of interest: First, it would improve discovery of the underlying mechanisms by not missing causative variables. Second, it would shed light on the molecular signature multiplicity phenomenon and how it affects the reproducibility of signatures.

Even though there are several well-developed algorithms for learning a Markov boundary (Aliferis et al., 2008a; Aliferis et al., 2003; Tsamardinos et al., 2003b), little research has been done in development of algorithms for identification of multiple Markov boundaries from the same dataset. Most notable advances in the field are described in the next subsection. In summary, there are currently no practical methods that can provably identify all Markov boundaries from the data without restrictions on the distribution.

The main focus of this thesis is development of a general theory and novel algorithms for identification of all Markov boundaries that exist in the underlying distribution. These algorithms can be applied to any type of data, independent of the distribution. In this thesis, I apply the novel algorithms to indentify the set of maximally predictive and non-redundant molecular signatures. I chose this application domain because of its importance and implications for biomedicine and personalized medicine. However, I would like to emphasize that the new algorithms by design can be applied to any type of data and problem domain, and I plan to explore this in the future. The experiments reported in the present thesis suggest that the new algorithms have excellent theoretical and empirical properties compared to the existing state-of-the-art methods.

The molecular signature multiplicity problem and its computational dissection

A *molecular signature* is a computational/mathematical model that predicts a phenotype of interest (e.g., diagnosis or outcome of treatment in human patients) from microarray gene expression or other high-throughput assay data inputs (Ramaswamy et al., 2003; Golub et al., 1999). *Multiplicity* is a special form of statistical instability in which different data analysis methods used on the same data, or different samples from the same population lead to *different*

but apparently maximally predictive signatures (Azuaje and Dopazo, 2005; Somorjai et al., 2003). This phenomenon has far-reaching implications for biological discovery and development of next generation patient diagnostics and personalized treatments. Multiplicity in the best case implies that generation of biological hypotheses (e.g., discovery of potential drug targets) is very hard even when signatures are maximally predictive of the phenotype since thousands of completely different signatures are equally consistent with the data. In the worst case this phenomenon entails that the produced signatures are not statistically generalizable to new cases, and thus not reliable enough for translation to clinical practice.

Some authors attribute signature multiplicity to the small sample size of typical microarray gene expression studies (Ein-Dor et al., 2006) and have conjectured that it leads to non-reproducible predictivity when the signatures are applied in independent data (Michiels et al., 2005). Related to the above, it has been suggested that building reproducible signatures requires thousands of observations (Ioannidis, 2005). Other authors proposed that the phenomenon of signature multiplicity is a byproduct of the complex regulatory connectivity of the underlying biological system leading to high predictive redundancy (Dougherty and Brun, 2006). This position implies that larger sample sizes may not reduce the number of maximally predictive molecular signatures. A third possible explanation of signature multiplicity is implicit in previously described artifacts of data pre-processing. For example, normalization may inflate correlations between genes, making some of them interchangeable for prediction of the phenotype (Qiu et al., 2005; Gold et al., 2005; Ploner et al., 2005).

A few computational methods have been recently introduced in an attempt to extract multiple signatures from the data aiming thus to provide practical tools for studying multiple maximally predictive signatures and the reasons for their existence. The methods encompass four algorithm families. The first family is *resampling-based signature extraction*. It operates by repeated application of a signature extraction algorithm to resampled data (e.g., via bootstrapping) (Roepman et al., 2006; Ein-Dor et al., 2005; Michiels et al., 2005). This family of

methods is based on the assumption that multiplicity is strictly a small sample phenomenon. To extract all true signatures an infinite number of resamplings is required in the worst-case. The second family is *iterative removal*, that is repeating signature extraction after removing from the data all genes that participate in the previously discovered molecular signatures (Natsoulis et al., 2005). This approach is agnostic as to what causes multiplicity. The third family is *stochastic gene selection* techniques (Peña et al., 2007; Li et al., 2001). The underlying premise of the method of (Peña et al., 2007) is that in a specific class of distributions every maximally predictive and non-redundant signature will be output by a randomized algorithm with non-zero probability (thus all such signatures will be output when the algorithm is applied an infinite number of times). Similarly, the method of (Li et al., 2001) will output all signatures discoverable by a genetic algorithm when it is allowed to evolve an infinite number of populations. The fourth family is *brute force exhaustive search* (Grate, 2005). This approach is also agnostic as to what causes multiplicity, and requires exponential time to the total number of genes, thus it is computationally infeasible for signatures with more than 2-3 genes (as almost all maximally predictive signatures are in practice).

The present work provides a theoretical framework based on Markov boundary induction that enables probabilistic modeling of multiple signatures and formally connects it with the causal graph of the data generating process (Guyon et al., 2007; Tsamardinos and Aliferis, 2003; Pearl, 2000; Pearl, 1988). The thesis introduces a provably correct algorithm (termed TIE*) that outputs all Markov boundaries (and by extension all maximally predictive and non-redundant signatures) independent of data distribution. I present experiments with real and resimulated microarray gene expression datasets as well as with artificial simulated data that verify the theoretical properties of TIE* and showcase its advantages over state-of-the-art methods. In particular, it is shown that TIE* having excellent sample and computational efficiency not only extracts many more maximally predictive and non-redundant signatures than all other available methods, but also that TIE* signatures reproduce in independent datasets whereas signatures produced by previous

methods are often not reproducible (i.e., they are overfitted). The theoretical and experimental results obtained in the present study also suggest that some of the previous hypotheses about the causes and implications of signature multiplicity have to be reevaluated.

Thesis organization

The remainder of this thesis is organized as follows: Chapter II presents a Markov boundary characterization of molecular signature multiplicity. Chapter III introduces the generative algorithm TIE* that outputs all Markov boundaries (and thus all maximally predictive and non-redundant signatures). Chapter IV provides a proof of correctness of the generative algorithm and proves admissibility of its instantiations. Chapter V describes results of empirical experiments with artificially simulated data where all Markov boundaries are known. Chapter VI presents results of empirical experiments with resimulated gene expression data that closely resembles real human gene expression data. Chapter VII presents results of an empirical evaluation of TIE* in real human microarray gene expression data. The thesis concludes with chapter VIII that reviews related methods from the field of statistics, discusses possible causes of the molecular signature multiplicity phenomenon, analyzes multiple signature extraction methods used in the thesis, provides directions for future research, and summarizes findings of this work. Supplementary materials are provided in the appendices.

CHAPTER II

MARKOV BOUNDARY CHARACTERIZATION OF MOLECULAR SIGNATURE MULTIPLICITY

Key definitions

Below I present three definitions that are essential for this thesis:

Definition of molecular signature: A molecular signature is a mathematical/ computational model (e.g., classifier or regression model) that predicts a phenotype of interest (e.g., diagnosis or response to treatment in human patients) given values of molecular variables (e.g., gene expression values).

Definition of maximally predictive molecular signature: A maximally predictive molecular signature is a molecular signature that maximizes predictivity of the phenotype relative to all other signatures that can be constructed from the given dataset.

Definition of maximally predictive and non-redundant molecular signature: A maximally predictive and non-redundant molecular signature based on variables **X** is a maximally predictive signature such that any signature based on a proper subset of variables in **X** is not maximally predictive.

The latter signatures that satisfy two critically desirable optimality properties (they are maximally predictive of the phenotype, and they do not contain predictively redundant genes) are the main focus of this thesis. Every suboptimal signature (i.e., one that is either not maximally predictive or contains redundant genes) can be discarded from consideration when studying multiplicity.

Markov boundary and its connection with the signature multiplicity phenomenon

Notation and basic definitions from the theory of learning graphical structures from data are states in Appendix A. Below I provide only definitions for key concepts that are required for understanding the theory.

First, I define the concept of Markov blanket and a related concept of Markov boundary.

Definition of Markov blanket: A *Markov blanket* **M** of the response variable $T \in \mathbf{V}$ in the joint probability distribution $\mathbb{P}$ over variables **V** is a set of variables conditioned on which all other variables are independent of *T*, i.e. for every $X \in (\mathbf{V} \setminus \mathbf{M} \setminus \{T\})$, $T \perp X \mid \mathbf{M}$.

Trivially, the set of all variables **V** excluding *T* is a Markov blanket of *T*. Also one can take a small Markov blanket and produce a larger one by adding arbitrary (predictively redundant) variables. Hence, only non-redundant Markov blankets are of interest.

Definition of Market boundary (non-redundant Markov blanket): If **M** is a Markov blanket of *T* and no proper subset of **M** satisfies the definition of Markov blanket of *T*, then **M** is called a *Markov boundary* (*non-redundant Markov blanket*) of *T*.

The following theorem states that variable sets that participate in the maximally predictive signatures of *T* are precisely the Markov blankets of *T*.

Theorem 1: If $\mathbb{M}$ is a performance metric that is maximized only when P(*T* | **V** \ {*T*}) is estimated accurately and $\mathbb{L}$ is a learning algorithm that can approximate any probability distribution, then **M** is a Markov blanket of *T* if and only if the learner's model induced using variables **M** is a maximally predictive signature of *T*.

Proof: First I prove that the learner's model induced using any Markov blanket of *T* is a maximally predictive signature of *T*. If **M** is Markov blanket of *T*, then by definition it leads to a maximally predictive signature of *T* because P(*T* | **M**) = P(*T* | **V** \ {*T*}) and this distribution can be perfectly approximated by $\mathbb{L}$, which implies that $\mathbb{M}$ will be maximized.

Now I prove that any maximally predictive signature of T is the learner's model induced using a Markov blanket of T. Assume that $\mathbf{X} \subseteq \mathbf{V} \setminus \{T\}$ is a set of variables used in the maximally predictive signature of T but it is not a Markov blanket of T. This implies that, $P(T \mid \mathbf{X}) \neq P(T \mid \mathbf{V} \setminus \{T\})$. By definition, $\mathbf{V} \setminus \{T\}$ is always a Markov blanket of T. By first part of the theorem, $\mathbf{V} \setminus \{T\}$ leads to a maximally predictive signature of T similarly to $\mathbf{X}$. Therefore, the following should hold: $P(T \mid \mathbf{X}) = P(T \mid \mathbf{V} \setminus \{T\})$. This contradicts the assumption that $\mathbf{X}$ is not a Markov blanket of T. Therefore, $\mathbf{X}$ is a Markov blanket of T. (Q.E.D.)

Since the notion of non-redundancy is defined in the same way for maximally predictive signatures and for Markov blankets, under the assumptions of Theorem 1 it follows that $\mathbf{M}$ is a Markov boundary of T if and only if the learner's model induced using variables $\mathbf{M}$ is a maximally predictive and non-redundant signature of T.

The next theorem provides a set of useful tools for theoretical analysis of probability distributions and proofs of correctness of Markov boundary algorithms. It is stated similarly to (Peña et al., 2007) and its proof is given in (Pearl, 1988).

Theorem 2: Let $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, and $\mathbf{W}$ be any[1] four subsets of variables from $\mathbf{V}$. The following four properties hold in any joint probability distribution $\mathbb{P}$ over variables $\mathbf{V}$:

- *Symmetry*: $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z} \Leftrightarrow \mathbf{Y} \perp \mathbf{X} \mid \mathbf{Z}$
- *Decomposition*: $\mathbf{X} \perp (\mathbf{Y} \cup \mathbf{W}) \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$ and $\mathbf{X} \perp \mathbf{W} \mid \mathbf{Z}$
- *Weak union*: $\mathbf{X} \perp (\mathbf{Y} \cup \mathbf{W}) \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp \mathbf{Y} \mid (\mathbf{Z} \cup \mathbf{W})$
- *Contraction*: $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$ and $\mathbf{X} \perp \mathbf{W} \mid (\mathbf{Z} \cup \mathbf{Y}) \Rightarrow \mathbf{X} \perp (\mathbf{Y} \cup \mathbf{W}) \mid \mathbf{Z}$

If $\mathbb{P}$ is strictly positive, then in addition to the above four properties a fifth property holds:

- *Intersection*: $\mathbf{X} \perp \mathbf{Y} \mid (\mathbf{Z} \cup \mathbf{W})$ and $\mathbf{X} \perp \mathbf{W} \mid (\mathbf{Z} \cup \mathbf{Y}) \Rightarrow \mathbf{X} \perp (\mathbf{Y} \cup \mathbf{W}) \mid \mathbf{Z}$

[1] Pearl originally provided this theorem for disjoint sets of variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ (Pearl, 1988). However, he mentioned that the disjoint requirement is made for the sake of clarity, and that the theorem can be extended to include overlapping subsets as well using an additional property $\mathbf{X} \perp \mathbf{Z} \mid \mathbf{Z}$ (denoted in this work as "*self-conditioning property*").

If $\mathbb{P}$ is faithful to $\mathbb{G}$, then $\mathbb{P}$ satisfies the above five properties and:

- *Composition:* $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$ and $\mathbf{X} \perp \mathbf{W} \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp (\mathbf{Y} \cup \mathbf{W}) \mid \mathbf{Z}$.

The following theorem states a sufficient assumption for the uniqueness of Markov boundaries.

Theorem 3: If a joint probability distribution $\mathbb{P}$ over variables $\mathbf{V}$ satisfies the intersection property, then for each $X \in \mathbf{V}$, there exists a unique Markov boundary of X (Pearl, 1988).

Since every joint probability distribution $\mathbb{P}$ that is faithful to $\mathbb{G}$ satisfies the intersection property (Theorem 2), then there is a unique Markov boundary in such distribution according to Theorem 3. However Theorem 3 does not say anything about distributions that do not satisfy the intersection property. I hypothesize that a joint probability distribution $\mathbb{P}$ that does not satisfy the intersection property can have multiple Markov boundaries.

The following two examples and a theorem provide graphical structures and related probability distributions where multiple Markov boundaries (and equivalently multiple maximally predictive and non-redundant signatures) exist. These examples also demonstrate that multiplicity of signatures exists even in large samples and thus it is not an exclusively small-sample phenomenon.

Example 2.1: Consider a joint probability distribution $\mathbb{P}$ described by a Bayesian network with graph $A \rightarrow B \rightarrow T$ where A, B, and T are binary random variables that take values {0, 1}. Given the local Markov condition, the joint probability distribution can be defined as follows: P(A=0) = 0.3, P(B=0 | A=1) = 1.0, P(B=1 | A=0) = 1.0, P(T=0 | B=1) = 0.2, P(T=0 | B=0) = 0.4. Two Markov boundaries of T exist in this distribution: {A} and {B}.

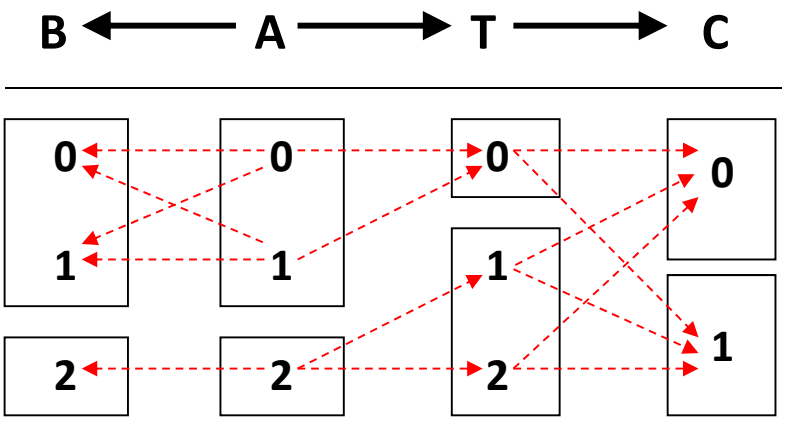
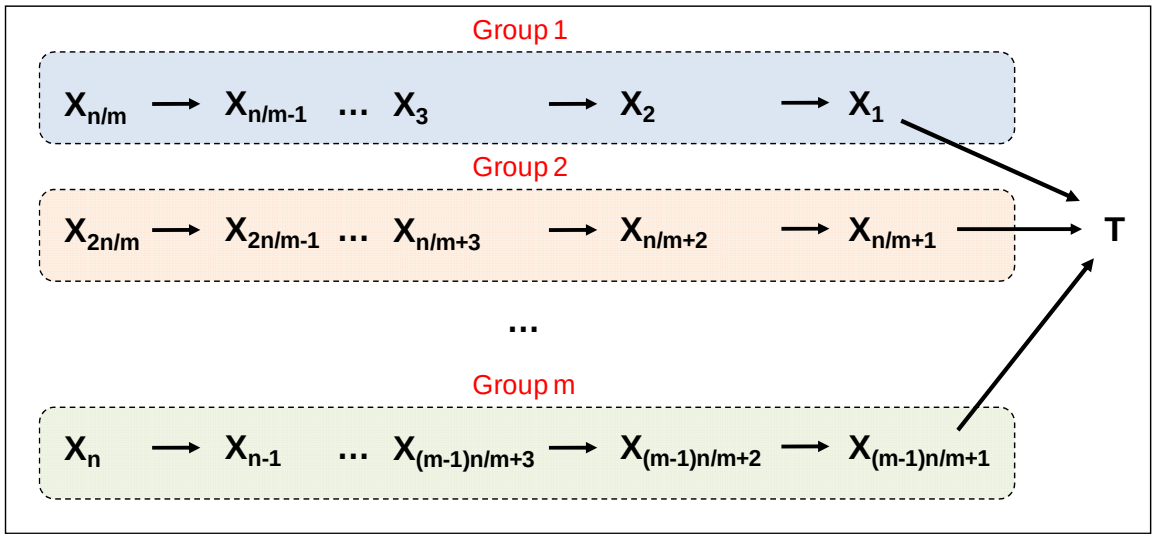


Figure 1: Graph of a Bayesian network with four variables (top) and constraints on its parameterization (bottom). Variables A, B, T take three values {0, 1, 2}, while variable C takes two values {0, 1}. Red dashed arrows denote nonzero conditional probabilities of each variable given its parents. For example, $P(T=0 \mid A=1) \neq 0$, while $P(T=0 \mid A=2) = 0$.

Example 2.2: Figure 1 shows a graph of a Bayesian network and constraints on its parameterization. The following hold in any joint probability distribution of a Bayesian network that satisfies the constraints in the figure:

- There exist two Markov boundaries of T: $\{A, C\}$ and $\{B, C\}$; Furthermore, $\{A, C\}$ and $\{B, C\}$ remain Markov boundaries of T even in infinite samples from that distribution;
- Variables A and B are not deterministically related, yet they convey individually the same information about T;
- If an algorithm selects only one Markov boundary of T (e.g., $\{B, C\}$), then there is danger to miss causative variables (i.e., parent A) and focus instead on confounded ones (i.e., B);
- The union of all Markov boundaries of T includes all variables located in the local neighborhood around T (i.e., A, C);
- In this example the intersection of all Markov boundaries of T contains only variables in the local neighborhood of T (i.e., C).

Also notice that the network in Figure 1 has very low connectivity (e.g., max in-degree = 1 and max out-degree = 2).



$P(T \mid X_1, X_{n/m+1},\ldots X_{(m-1)n/m+1})$	$(X_1=0, X_{n/m+1}=0,\ldots X_{(m-1)n/m+1}=0)$	$(X_1=0, X_{n/m+1}=0,\ldots X_{(m-1)n/m+1}=1)$	…	$(X_1=1, X_{n/m+1}=1,\ldots X_{(m-1)n/m+1}=1)$
$T=0$	0.2	0.8		0.2
$T=1$	0.8	0.2		0.8

For any pair of variables X_j and X_k belonging to the same group i:

$P(X_j \mid X_k)$	$X_k=0$	$X_k=1$
$X_j=0$	1.0	0.0
$X_j=1$	0.0	1.0

Figure 2: Graph of a Bayesian network used to demonstrate that the number of Markov boundaries can be exponential to the number of variables in the network. The network parameterization of is provided below the graph. The response variable is T. All variables take values $\{0, 1\}$. All variables X_i in each group provide exactly the same information about T.

Theorem 4: The number of Markov boundaries can grow exponentially with the number of variables.

Proof: I prove this theorem constructively by providing an example network and probability distribution where the number of Markov boundaries grows exponentially with the number of variables. Consider a Bayesian network shown in Figure 2. It involves $n+1$ binary variables: $X_1, X_2, ..., X_n$, and a response variable T. Variables X_i ($i = 1,\ldots,n$) can be divided into m groups such that any two variables in a group contain exactly the same information about T. Since

there are n/m variables in each group, the total number of Markov boundaries is $(n/m)^m$. Now assume that $m = kn$, where $k < 1$. Then the total number of Markov boundaries is $(1/k)^{kn}$. Since $1/k > 1$ and $kn = O(n)$, it follows that the number of Markov boundaries grows exponentially with the number of variables in this example. (Q.E.D.)

The above discussion is concerned with the large sample case. In practice, one deals with small samples where statistical inferences have to be made about large sample predictivity and redundancy. This creates an additional source of error and concomitant multiplicity as illustrated in chapter VI (experiment 1) and Appendix J.

A fundamental assumption for the analysis of signatures

To simplify analysis, and without loss of generality, from now on instead of considering all possible signatures derivable from a given dataset (via a potentially infinite variety of classifier algorithms), I only consider the signatures that have maximal predictivity for the phenotypic response variable *relative to the genes (variables) contained in each signature*. In other words, I exclude from consideration signatures that do not utilize all predictive information contained in their genes. This allows to study signature classes by reference only to the genes contained in each class. Specifically, for a gene set **X** there can be an infinite number of classifiers that achieve maximal predictivity for the phenotype relative to the information contained in **X**. Thus, for the remainder of this thesis when I say "signature **X**" I refer to one of these predictively equivalent classifiers. This reduction is justified whenever the classifiers used can learn the minimum error decision function[2] given sufficient sample. Most practical classifiers employed in this domain as well as classifiers used in the present experiments (SVMs) satisfy the above requirement either on theoretical (Shawe-Taylor and Cristianini, 2004; Hammer and

[2] For a given set of genes **S**, the minimal error decision function minimizes the error of predicting the phenotypic variable T given **S** over all possible decision functions.

Gersmann, 2003) and/or empirical grounds (Statnikov et al., 2008; Statnikov et al., 2005; Furey et al., 2000).

Figure 3 provides an example of a dataset with two genes X_1 and X_2 and a phenotypic response variable T. There are two classes of signatures: ones that have maximal predictivity of the phenotype relative to their genes (e.g., signatures $\mathbf{S}_3$, $\mathbf{S}_4$, $\mathbf{S}_5$ that predict T without errors) and ones with worse predictivity (e.g., signatures $\mathbf{S}_1$, $\mathbf{S}_2$). Each of the classes contains an infinite number of signatures. When I say "signature $\{X_1, X_2\}$" in this thesis, I mean one of the predictively equivalent classifiers with maximal predictivity of the phenotype, e.g. $\mathbf{S}_3$, or $\mathbf{S}_4$, or $\mathbf{S}_5$,, etc.

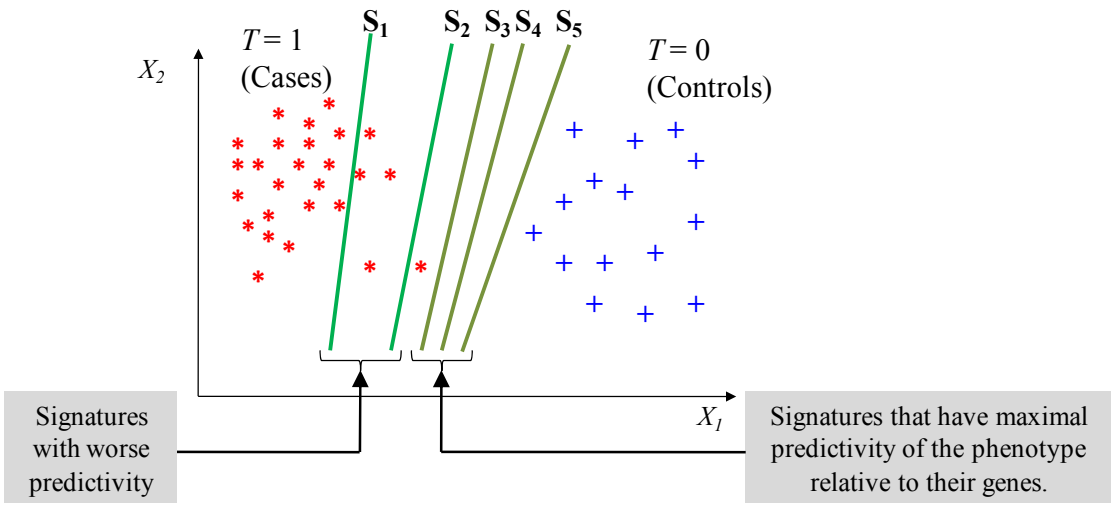


Figure 3: Graph of an example dataset with two genes X_1 and X_2 and a phenotypic response variable T. Two classes of signatures exist in the data: signatures with maximal predictivity of the phenotype relative to their genes and ones with worse predictivity. There is an infinite number of signatures in each class.

CHAPTER III

NOVEL ALGORITHM

The TIE* algorithm pseudocode is provided in Figure 4. It is a generative algorithm that is instantiated differently for different distributions (an example of instantiated TIE* algorithm for gene expression data analysis is provided in Figure 5). On input the generative algorithm receives (i) a dataset $\mathbb{D}$ (a sample of distribution $\mathbb{P}$) for variables **V**, including a response variable *T*; (ii) a Markov boundary algorithm $\mathbb{X}$; (iii) a strategy $\mathbb{Y}$ to generate subsets of variables that have to be removed from **V** to indentify new Markov boundaries of *T*; and (iv) a criterion $\mathbb{Z}$ to verify Markov boundaries of *T*. The input components $\mathbb{X}$, $\mathbb{Y}$, $\mathbb{Z}$ are selected to be suitable for the distribution in hand and should satisfy admissibility rules stated in Figure 6 for correctness of the algorithm. The algorithm outputs all Markov boundaries (i.e., all maximally predictive and non-redundant signatures) of *T*.

In line 1, TIE* uses a Markov boundary algorithm $\mathbb{X}$ to learn a Markov boundary **M** of *T* from data $\mathbb{D}$ for variables **V** (i.e., in the *original* distribution). Then **M** is output in line 2. In line 4, the algorithm uses a strategy $\mathbb{Y}$ to generate a subset **G** whose removal may lead to identification of a new Markov boundary of *T*. Next, in line 5 the Markov boundary algorithm $\mathbb{X}$ is applied to a version of the dataset $\mathbb{D}$ in which a subset of variables **G** has been removed (I refer to this as *embedded* distribution), resulting in a Markov boundary $\mathbf{M}_{new}$ in the embedded distribution. If $\mathbf{M}_{new}$ is also a Markov boundary of *T* in the original distribution according to criterion $\mathbb{Z}$, then $\mathbf{M}_{new}$ is output (line 6). The loop in lines 3-7 is repeated until all subsets **G** generated by strategy $\mathbb{Y}$ have been considered.

Generative algorithm **TIE***

Inputs:

- dataset $\mathbb{D}$ (a sample of distribution $\mathbb{P}$) for variables **V**, including a response variable *T*;
- Markov boundary algorithm $\mathbb{X}$;
- strategy $\mathbb{Y}$ to generate subsets of variables that have to be removed to identify new Markov boundaries of *T*;
- criterion $\mathbb{Z}$ to verify Markov boundaries of *T*.

Output: all Markov boundaries (i.e., maximally predictive and non-redundant signatures) of *T*.

1. Use algorithm $\mathbb{X}$ to learn a Markov boundary **M** of *T* from data $\mathbb{D}$ for variables **V** (i.e., in the *original* distribution)
2. Output **M**
3. Repeat
4. Use strategy $\mathbb{Y}$ to generate a subset of variables **G** whose removal may lead to identification of a new Markov boundary of *T*
5. Use algorithm $\mathbb{X}$ to learn a Markov boundary $\mathbf{M}_{new}$ of *T* from data $\mathbb{D}$ for variables **V** \ **G** (i.e., in the *embedded* distribution)
6. If $\mathbf{M}_{new}$ is a Markov boundary of *T* in the original distribution according to criterion $\mathbb{Z}$, output $\mathbf{M}_{new}$
7. Until all subsets **G** generated by strategy $\mathbb{Y}$ have been considered.

Figure 4: TIE* generative algorithm.

An example of instantiated algorithm **TIE*** for gene expression data analysis

Inputs: dataset $\mathbb{D}$ (a sample of distribution $\mathbb{P}$) for variables **V**, including a response variable *T*;

Output: all Markov boundaries (i.e., maximally predictive and non-redundant signatures) of *T*.

1. Use algorithm HITON-PC to learn a Markov boundary **M** of *T* from data $\mathbb{D}$ for variables **V** (i.e., in the *original* distribution)
2. Output **M**
3. Repeat
4. Generate the smallest subset **G** of the so far discovered Markov boundaries of *T* such that (i) it was not considered in the previous iteration of the algorithm, and (ii) it does not properly include any subset that was generated in the previous iteration of the algorithm when $\mathbf{M}_{new}$ was found not to be a Markov boundary of *T*
5. Use algorithm HITON-PC to learn a Markov boundary $\mathbf{M}_{new}$ of *T* from data $\mathbb{D}$ for variables **V** \ **G** (i.e., in the *embedded* distribution)
6. If the holdout validation estimate of predictivity of *T* for the SVM classifier model induced from data $\mathbb{D}$ using variables $\mathbf{M}_{new}$ is statistically indistinguishable from the respective predictivity estimate for variables **M**, then $\mathbf{M}_{new}$ is a Markov boundary of *T* in the original distribution and it is output by the algorithm
7. Until no subset **G** can be generated in line 4.

Figure 5: An example of instantiated TIE* algorithm for gene expression data analysis.

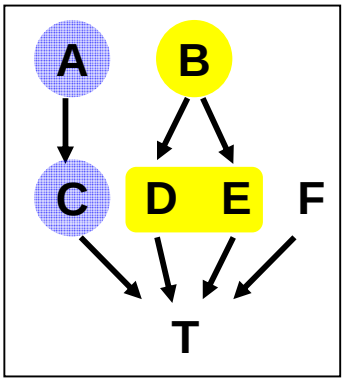
Admissibility rules for inputs $\mathbb{X}$, $\mathbb{Y}$, $\mathbb{Z}$ of the TIE* algorithm

I. The Markov boundary algorithm $\mathbb{X}$ correctly identifies a Markov boundary of T both in the original distribution (i.e., for variables **V**) and in every embedded distribution that is obtained by removing from **V** a subset of variables generated by $\mathbb{Y}$.
II. The strategy $\mathbb{Y}$ to generate subsets of variables is complete, i.e. it will generate every subset **G** that is needed to be removed from **V** to identify every Markov boundary of T.
III. The criterion $\mathbb{Z}$ can correctly identify that $\mathbf{M}_{new}$ is a Markov boundary of T in the original distribution.

Figure 6: Admissibility rules for inputs $\mathbb{X}$, $\mathbb{Y}$, $\mathbb{Z}$ of the TIE* algorithm.

Consider running TIE* algorithm on data $\mathbb{D}$ generated from the example Bayesian network shown in Figure 7. The response variable T is directly caused by C, D, E, and F. The underlying distribution is such that variables A and C contain exactly the same information about T; likewise two variables $\{D, E\}$ jointly and a single variable B contain exactly the same information about T. In line 1 of TIE* (Figure 4), a Markov boundary algorithm $\mathbb{X}$ is applied to learn a Markov boundary of T: $\mathbf{M} = \{A, B, F\}$. Then **M** is output in line 2. In line 4, the strategy $\mathbb{Y}$ generates a subset $\mathbf{G} = \{F\}$ whose removal may lead to identification of a new Markov boundary of T. Then in line 5 the Markov boundary algorithm $\mathbb{X}$ is run on data $\mathbb{D}$ for all variables but F (i.e., in the embedded distribution). This yields a Markov boundary of T in the embedded distribution $\mathbf{M}_{new} = \{A, B\}$. The criterion $\mathbb{Z}$ in line 6 does not confirm that $\mathbf{M}_{new}$ is also Markov boundary of T in the original distribution; thus $\mathbf{M}_{new}$ is not output. The loop is run again. In line 4 the strategy $\mathbb{Y}$ generates another subset $\mathbf{G} = \{A\}$. The Markov boundary algorithm $\mathbb{X}$ in line 5 yields a Markov boundary of T in the embedded distribution $\mathbf{M}_{new} = \{C, B, F\}$. The criterion Z in line 6 confirms that $\mathbf{M}_{new}$ is also a Markov boundary in the original distribution, thus it is output. Similarly, when the Markov boundary algorithm $\mathbb{X}$ is run on data $\mathbb{D}$ for all variables but $\mathbf{G} = \{B\}$ or $\mathbf{G} = \{A, B\}$, two more Markov boundaries of T in the original distribution, $\{A, D, E, F\}$ or $\{C, D, E, F\}$, respectively, are found and output. The algorithm terminates shortly. In total, four

Markov boundaries of T are output by the algorithm: $\{A, B, F\}$, $\{C, B, F\}$, $\{A, D, E, F\}$ and $\{C, D, E, F\}$. These are exactly all Markov boundaries of T that exist in this distribution.



P(A)	
$A = 0$	0.6
$A = 1$	0.4

P(B)	
$B = 0$	0.3
$B = 1$	0.2
$B = 2$	0.3
$B = 3$	0.2

P($C \mid A$)	$A = 0$	$A = 1$
$C = 0$	0.0	1.0
$C = 1$	1.0	0.0

P(F)	
$F = 0$	0.3
$F = 1$	0.7

P($D \mid B$)	$B = 0$	$B = 1$	$B = 2$	$B = 3$
$D = 0$	1.0	1.0	0.0	0.0
$D = 1$	0.0	0.0	1.0	1.0

P($E \mid B$)	$B = 0$	$B = 1$	$B = 2$	$B = 3$
$E = 0$	1.0	0.0	1.0	0.0
$E = 1$	0.0	1.0	0.0	1.0

P($T \mid C, D, E, F$)	(C=0, D=0, E=0, F=0)	(C=0, D=0, E=0, F=1)	(C=0, D=0, E=1, F=0)	…	(C=1, D=1, E=1, F=1)
$T = 0$	0.9	0.1	0.9	…	0.1
$T = 1$	0.1	0.9	0.1		0.9

Figure 7: Graph of a Bayesian network used to trace the TIE* algorithm. The network parameterization is provided below the graph. The response variable is T. All variables take values $\{0, 1\}$ except for B that takes values $\{0, 1, 2, 3\}$. Variables A and C contain exactly the same information about T and are highlighted with the same color. Likewise, two variables $\{D, E\}$ jointly and a single variables B contain exactly the same information about T and thus are also highlighted with the same color.

CHAPTER IV

THEORETICAL ANALYSIS OF THE NOVEL ALGORITHM AND ITS ADMISSIBLE INSTANTIATIONS

Proof of correctness of the generative algorithm TIE*

Theorem 5: The generative algorithm TIE* outputs all and only Markov boundaries of T if the input components $\mathbb{X}$, $\mathbb{Y}$, $\mathbb{Z}$ are admissible.

Proof: TIE* will trivially output only Markov boundaries of T when the input components $\mathbb{X}$ and $\mathbb{Z}$ are admissible (Figure 6). Assume that there exists a Markov boundary **W** that is not output by TIE*. Also assume that **W** does not overlap with any other Markov boundary output by TIE* (the proof is similar if **W** has such an overlap). Because of admissibility of input components $\mathbb{X}$ and $\mathbb{Z}$ (Figure 6), $\mathbf{M}_{new} = \mathbf{W}$ was never identified in line 5 of the algorithm. This can happen if and only if $T \perp \mathbf{W} \mid \mathbf{M}_i$ where $\mathbf{M}_i$ is some Markov boundary that was previously discovered by TIE* (either in line 1 or 5). However, because of admissibility of input component $\mathbb{Y}$ (Figure 6) in some iteration of the algorithm in line 4 the subset $\mathbf{G} = \mathbf{M}_i$ (and similarly all other subsets that render **W** independent of T) will be generated and removed from the dataset in line 5. Thus **W** will be discovered in line 5 and output in line 6. Therefore, a contradiction is reached, and TIE* would never miss Markov boundaries. (Q.E.D.)

I also note that the above proof of correctness holds when the admissibility criterion for Markov boundary algorithm $\mathbb{X}$ is relaxed in such a way that $\mathbb{X}$ may not correctly identify a Markov boundary $\mathbf{M}_{new}$ in the embedded distribution when there is no Markov boundary in the embedded distribution that is also a Markov boundary in the original distribution.

Admissibility analysis of the Markov boundary algorithms

First, I prove admissibility of the Markov boundary algorithm IAMB (Tsamardinos and Aliferis, 2003; Tsamardinos et al., 2003a) that is described in Figure 8. To do this, I need to define a relaxed version of the composition property:

Definition of local composition property with respect to a variable: Let **X**, **Y**, **Z** be any three subsets of variables from **V**. The joint probability distribution $\mathbb{P}$ over variables **V** satisfies the local composition property with respect to T if $T \perp \mathbf{X} \mid \mathbf{Z}$ and $T \perp \mathbf{Y} \mid \mathbf{Z} \Rightarrow T \perp (\mathbf{X} \cup \mathbf{Y}) \mid \mathbf{Z}$.

Originally the IAMB algorithm was shown to be correct (i.e., that it identifies a Markov boundary) if the joint probability distribution $\mathbb{P}$ is DAG-faithful to $\mathbb{G}$ (Tsamardinos and Aliferis, 2003; Tsamardinos et al., 2003a). The following theorem originally proven in (Peña et al., 2007)[3] shows that IAMB is correct when only the local composition property with respect to T holds.

Algorithm **IAMB**

Input: dataset $\mathbb{D}$ (a sample of distribution $\mathbb{P}$) for variables **V**, including a response variable T.

Output: a Markov boundary **M** of T.

Phase I: Forward

1. Initialize **M** with an empty set
2. Repeat
3. $Y \leftarrow \text{argmax}_{X \in (\mathbf{V} \setminus \mathbf{M} \setminus \{T\})} \text{Association}(T, X \mid \mathbf{M})$
4. If $T \not\perp Y \mid \mathbf{M}$ then
5. $\mathbf{M} \leftarrow \mathbf{M} \cup \{Y\}$
6. Until **M** does not change

Phase II: Backward

7. For each $X \in \mathbf{M}$
8. If $T \not\perp X \mid (\mathbf{M} \setminus \{X\})$ then
9. $\mathbf{M} \leftarrow \mathbf{M} \setminus \{X\}$
10. End
11. Output **M**

Figure 8: IAMB algorithm.

[3] Peña et al. originally proved correctness of IAMB when the (global) composition property holds.

Theorem 6: IAMB outputs a Markov boundary of *T* if the joint probability distribution $\mathbb{P}$ satisfies the local composition property with respect to *T*. (The proof is given in Appendix C).

The proof of admissibility of IAMB follows below.

Theorem 7: IAMB is admissible Markov boundary algorithm for TIE* if the joint probability distribution $\mathbb{P}$ satisfies the local composition property with respect to *T*.

Proof: Since (i) all variables from each embedded distribution belong to the original one and (ii) the joint probability distribution of variables in each embedded distribution is the same as marginal in the original one, the local composition property with respect to *T* also holds in each embedded distribution. Therefore according to Theorem 6, IAMB will correctly identify a Markov boundary in every embedded distribution. (Q.E.D.)

Next, I prove admissibility of the Markov boundary algorithm HITON-PC (Aliferis et al., 2008a; Aliferis et al., 2003) that is described in Figure 9. Originally this algorithm was shown to correctly identify a set of parents and children of *T* if the joint probability distribution $\mathbb{P}$ is DAG-faithful to $\mathbb{G}$ and the so-called "symmetry correction" is not required (Aliferis et al., 2008a). Below I prove correctness of this algorithm for identification of Markov boundaries when the intersection property may be violated. This proof requires revisiting the assumption of faithfulness and introducing several new definitions, see Appendix B.

Theorem 8: HITON-PC outputs a Markov boundary of *T* if (i) the joint probability distribution $\mathbb{P}$ and directed or ancestral graph $\mathbb{G}$ are locally adjacency faithful with respect to *T* with the exception of violations of the intersection property; (ii) $\mathbb{P}$ satisfies the global Markov condition for $\mathbb{G}$; (iii) the set of vertices adjacent with *T* in $\mathbb{G}$ is a Markov blanket of *T*. (The proof is given in Appendix C).

It is worthwhile to note that the so-called "symmetry correction" is not needed for correctness of HITON-PC because the condition (iii) of the theorem subsumes it.

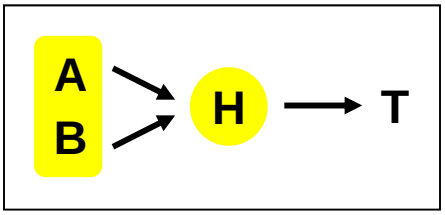
Algorithm **HITON-PC** (without "symmetry correction")

Input: dataset $\mathbb{D}$ (a sample of distribution $\mathbb{P}$) for variables **V**, including a response variable T.

Output: a Markov boundary **M** of T.

1. Initialize **M** with an empty set
2. Initialize the set of eligible variables **E** ← **V** \ $\{T\}$
3. Sort in descending order the variables in **E** according to their pairwise association with response variable T
4. Remove from **E** all variables X with zero association with T, i.e. when $T \perp X$
5. Repeat
6. X ← first variable in **E**
7. Add X to **M** and remove it from **E**
8. If $\exists\, \mathbf{Z} \subseteq \mathbf{M} \setminus \{X\}$, such that $T \perp X \mid \mathbf{Z}$, remove X from **M**
9. Until **E** is empty
10. For each $X \in \mathbf{M}$
11. If $\exists\, \mathbf{Z} \subseteq \mathbf{M} \setminus \{X\}$, such that $T \perp X \mid \mathbf{Z}$, remove X from **M**
12. Output **M**

Figure 9: HITON-PC algorithm (without "symmetry correction").



P(A)	
$A = 0$	0.5
$A = 1$	0.5

P(B)	
$B = 0$	0.5
$B = 1$	0.5

P($H \mid A, B$)	(A=0, B=0)	(A=0, B=1)	(A=1, B=0)	(A=1, B=1)
$H = 0$	1	0	0	1
$H = 1$	0	1	1	0

P($T \mid H$)	$H = 0$	$H = 1$
$T = 0$	0.3	0.8
$T = 1$	0.7	0.2

Figure 10: Graph of a Bayesian network used to motivate a more restrictive faithfulness assumption for admissibility of HITON-PC in the TIE* algorithm. The network parameterization is provided below the graph. The response variable is T. All variables take values {0, 1}. Two variables {A, B} jointly and a single variables H contain exactly the same information about T and thus are also highlighted with the same color.

Before I actually prove admissibility of the Markov boundary algorithm HITON-PC for TIE*, I will demonstrate why HITON-PC is not admissible for TIE* under the assumptions of Theorem 8 and why more stringent assumptions are needed.

One of the assumptions for correctness of HITON-PC is that $\mathbb{P}$ and $\mathbb{G}$ are locally adjacency faithful with respect to T with the exception of violations of the intersection property. Consider a Bayesian network shown in Figure 10. It follows that $\mathbb{P}$ and $\mathbb{G}$ are locally adjacency faithful with respect to T. However, when the variable H is removed, the resulting embedded joint probability distribution defined over $\{T, A, B\}$ will not be locally adjacency faithful with respect to T to any directed or ancestral graph. Therefore, HITON-PC would not discover that $\{A, B\}$ is a Markov boundary of T in the embedded distribution.

Another assumption for correctness of HITON-PC is that the set of vertices adjacent with T in $\mathbb{G}$ is a Markov blanket of T. Consider a Bayesian network specified by the graph $\mathbb{G}$: $T \rightarrow H \rightarrow A \leftarrow B$ and the joint probability distribution $\mathbb{P}$ that is DAG-faithful to $\mathbb{G}$. When the variable H is removed, the resulting embedded joint probability distribution defined over $\{T, A, B\}$ will be DAG-faithful to the graph $T \rightarrow A \leftarrow B$. However, notice that $\{A\}$ (the set of vertices adjacent with T) is not a Markov blanket of T, because the variable B is also present in the Markov blanket of T. In such case, HITON-PC would incorrectly discover that $\{A\}$ is a Markov boundary of T in the embedded distribution.

The proof of admissibility of HITON-PC follows below.

Theorem 9: HITON-PC is admissible Markov boundary algorithm for TIE* if (i) the joint probability distribution $\mathbb{P}$ and DAG $\mathbb{G}$ are locally path faithful with respect to T with the exception of violations of the intersection property; (ii) $\mathbb{P}$ satisfies the global Markov condition for $\mathbb{G}$; and (iii) the set of vertices adjacent with T both in DAG $\mathbb{G}$ and corresponding MAG $\mathbb{G}^*$ of the embedded distribution is a Markov blanket of T (in the respective distribution).

Proof: First I prove that HITON-PC correctly identifies a Markov boundary of T in the original distribution (i.e., for variables $\mathbf{V}$). To do this, I need to demonstrate that assumptions of Theorem 8 are satisfied. Since local path faithfulness with respect to T implies local adjacency faithfulness with respect to T (with the exception of violations of the intersection property), and the other two assumptions (ii) and (iii) are same for both Theorems 8 and 9, HITON-PC correctly identifies a Markov boundary of T in the original distribution.

Now I need to prove that HITON-PC also correctly identifies a Markov boundary of T in the embedded distribution, after removing an arbitrary subset of variables from $\mathbf{V}$. Again, I need to demonstrate that assumptions of Theorem 8 are satisfied for every embedded distribution. Consider an embedded distribution defined over variables $\mathbf{V}^* = \mathbf{V} \setminus \mathbf{S}$ (where $\mathbf{S}$ is a subset of $\mathbf{V}$ that is hidden/removed) with the joint probability distribution $\mathbb{P}^*=\mathbb{P}(\mathbf{V}^*)$ and graph $\mathbb{G}^*=<\mathbf{V}^*, \mathbb{E}^*>$. Given a DAG $\mathbb{G}$ of the original distribution and the subset of hidden variables $\mathbf{S}$, $\mathbb{G}^*$ is defined as follows: for every pair of variables X and Y, put an edge between them if and only if they are not d-separated in $\mathbb{G}$ by any subset of variables $\mathbf{V}^* \setminus \{X, Y\}$; the arrowhead of the edge is pointed at X (Y) if it is not an ancestor of Y (X) in $\mathbb{G}$. $\mathbb{G}^*$ is a maximal ancestral graph (MAG) and has the property that for any two non-adjacent vertices there is a set of vertices that m-separates them (Zhang and Spirtes, 2005).

- Assumption (i): Assume that $\mathbb{P}^*$ and $\mathbb{G}^*$ are not locally adjacency faithful with respect to T, excluding violations of the intersection property. In other words, there is a variable Y that is adjacent with T in $\mathbb{G}^*$ and T can be rendered independent of Y given some subset of variables $\mathbf{V}^* \setminus \{T, Y\}$. Since Y is adjacent with T in $\mathbb{G}^*$, it should be connected to T by a path in $\mathbb{G}$ that does not contain any colliders. Assume that this path is $T - X_1 - X_2 - \ldots - X_N - Y$ and there are no other paths without colliders that connect T and Y in $\mathbb{G}$ (the proof is similar when there are multiple paths). The local path faithfulness with respect to T in the original distribution implies that T cannot be rendered independent of Y given any

subset of $\mathbf{V} \setminus \{X_1, \ldots, X_N\}$. If $\mathbf{V}^* \cap \{X_1, \ldots, X_N\} = \varnothing$, then a contradiction of the assumption that T can be rendered independent of Y given some subset of variables $\mathbf{V}^* \setminus \{T, Y\}$ is reached. Otherwise when $\mathbf{V}^* \cap \{X_1, \ldots, X_N\} \neq \varnothing$, a contradiction of the assumption that Y is adjacent with T in $\mathbb{G}^*$ is reached, because Y will be adjacent with X_i and not with Y (where i is the minimal index of the variable X_i that belongs to $\mathbf{V}^* \cap \{X_1, \ldots, X_N\}$). Therefore, $\mathbb{P}^*$ and $\mathbb{G}^*$ will be locally adjacency faithful with respect to T with the exception of violations of the intersection property.

- Assumption (ii): The global Markov condition holds in the embedded distribution since all variables from the embedded distribution belong to the original one and the joint probability distribution of variables in the embedded distribution is the same as marginal in the original one.
- Assumption (iii) is satisfied by design because the set of vertices adjacent with T in MAG $\mathbb{G}^*$ is a Markov blanket of T in the embedded distribution.

Since all assumptions of Theorem 8 are satisfied, HITON-PC correctly identifies a Markov boundary of T in every embedded distribution. Thus, HITON-PC is an admissible algorithm for TIE*. (Q.E.D.)

Admissibility analysis of the criteria to verify Markov boundaries

Theorem 10: Criterion *Independence* to verify Markov boundaries (Figure 11) is admissible for TIE*.

Proof: Consider that there exists a set of variables $\mathbf{M}_{new} \subseteq \mathbf{V} \setminus \{T\}$ such that $T \perp \mathbf{M} \mid \mathbf{M}_{new}$. Since $\mathbf{M}$ is a Markov boundary of T in the original distribution, it is also a Markov blanket of T in the original distribution. By definition of the Markov blanket, $T \perp (\mathbf{V} \setminus \mathbf{M} \setminus \{T\}) \mid \mathbf{M}$. By the self-conditioning property, it follows that $T \perp (\mathbf{V} \setminus \{T\}) \mid \mathbf{M}$. Since $(\mathbf{V} \setminus \{T\}) = (\mathbf{V} \setminus \{T\}) \cup \mathbf{M}_{new}$ and according to the weak union property,

$T \perp (\mathbf{V} \setminus \{T\} \setminus \mathbf{M}_{new}) \mid (\mathbf{M} \cup \mathbf{M}_{new})$. By the self-conditioning property, it follows that $T \perp (\mathbf{V} \setminus \{T\}) \mid (\mathbf{M} \cup \mathbf{M}_{new})$. Since $T \perp \mathbf{M} \mid \mathbf{M}_{new}$ and $T \perp (\mathbf{V} \setminus \{T\}) \mid (\mathbf{M} \cup \mathbf{M}_{new})$, the contraction property implies that $T \perp ((\mathbf{V} \setminus \{T\}) \cup \mathbf{M}) \mid \mathbf{M}_{new}$. Since $(\mathbf{V} \setminus \{T\}) = (\mathbf{V} \setminus \{T\}) \cup \mathbf{M}$, it follows that $T \perp (\mathbf{V} \setminus \{T\}) \mid \mathbf{M}_{new}$. By the decomposition property this implies that $\mathbf{M}_{new}$ is a Markov blanket of T in the original distribution. Since $\mathbf{M}_{new}$ is a Markov boundary of T in the embedded distribution and it is a Markov blanket of T in the original distribution, it is also a Markov boundary of T in the original distribution. (Q.E.D.)

The above proof implicitly assumes correctness of statistical decisions about independence. In practice, this assumption may be violated when the sample size is small or the sampling of the dataset $\mathbb{D}$ is not i.i.d.

Criterion *Independence* to verify Markov boundaries

Inputs:

- dataset $\mathbb{D}$ (a sample of distribution $\mathbb{P}$) for variables **V**, including a response variable T;
- Markov boundary **M** of T in the original distribution;
- Markov boundary $\mathbf{M}_{new}$ of T in the embedded distribution;

Output:

- TRUE if $\mathbf{M}_{new}$ is a Markov boundary of T in the original distribution;
- FALSE if $\mathbf{M}_{new}$ is a not a Markov blanket of T in the original distribution.

If $T \perp \mathbf{M} \mid \mathbf{M}_{new}$, output TRUE; otherwise output FALSE.

Figure 11: Criterion *Independence* to verify Markov boundaries.

Theorem 11: Criterion *Predictivity* to verify Markov boundaries (Figure 12) is admissible for TIE* if the following conditions hold:

- learning algorithm $\mathbb{L}$ can accurately approximate any probability distribution;
- performance metric $\mathbb{M}$ is maximized only when $P(T \mid \mathbf{V} \setminus \{T\})$ is estimated accurately;
- performance estimator $\mathbb{E}$ is unbiased;
- procedure $\mathbb{C}$ to compare performance estimates of metric $\mathbb{M}$ has negligible error.

Criterion *Predictivity* to verify Markov boundaries

Inputs:

- dataset $\mathbb{D}$ (a sample of distribution $\mathbb{P}$) for variables **V**, including a response variable *T*;
- Markov boundary **M** of *T* in the original distribution;
- Markov boundary $\mathbf{M}_{new}$ of *T* in the embedded distribution;
- learning algorithm $\mathbb{L}$ (to build a prediction model for *T* given data $\mathbb{D}$ for some subset of variables **V**);
- performance metric $\mathbb{M}$ (to assess the prediction model obtained by $\mathbb{L}$; larger values of this performance metric correspond to better predictivity of the model);
- unbiased performance estimator $\mathbb{E}$ (to estimate metric $\mathbb{M}$ for prediction model obtained by $\mathbb{L}$ in data $\mathbb{D}$);
- statistical hypothesis test or another formal criterion $\mathbb{C}$ (to compare performance estimates of $\mathbb{M}$).

Output:

- TRUE if $\mathbf{M}_{new}$ is a Markov boundary of *T* in the original distribution;
- FALSE if $\mathbf{M}_{new}$ is a not a Markov blanket of *T* in the original distribution.

1. Apply performance estimator $\mathbb{E}$ to compute estimate $\hat{M}^1$ of performance metric $\mathbb{M}$ for prediction model obtained by $\mathbb{L}$ in data $\mathbb{D}$ using variables **M**
2. Apply performance estimator $\mathbb{E}$ to compute estimate $\hat{M}^2$ of performance metric $\mathbb{M}$ for prediction model obtained by $\mathbb{L}$ in data $\mathbb{D}$ using variables $\mathbf{M}_{new}$
3. If the hypothesis $\hat{M}^1 > \hat{M}^2$ can be rejected according to criterion $\mathbb{C}$, output TRUE; otherwise output FALSE.

Figure 12: Criterion *Predictivity* to verify Markov boundaries.

Proof: The proof that this criterion can identify whether $\mathbf{M}_{new}$ is a Markov blanket of *T* in the original distribution or not follows from Theorem 1. Since $\mathbf{M}_{new}$ is a Markov boundary of *T* in the embedded distribution and if it is a Markov blanket of *T* in the original distribution, it is also a Markov boundary of *T* in the original distribution. (Q.E.D.)

Admissibility analysis of the strategies to generate subsets of variables to that have to be removed to identify new Markov boundaries

Theorem 12: Strategies *IncLex*, *IncMinAssoc*, and *IncMaxAssoc* to generate subsets of variables that have to be removed from **V** to identify new Markov boundaries of *T* (Figure 13) are admissible for TIE*.

Strategies *IncLex, IncMinAssoc*, and *IncMaxAssoc* to generate subsets of variables that have to be removed from **V** to identify new Markov boundaries of *T*

Inputs:

- dataset $\mathbb{D}$ (a sample of distribution $\mathbb{P}$) for variables **V**, including a response variable *T*;
- Markov boundaries $\mathbf{M}_1,...,\mathbf{M}_n$ of *T* (in the original distribution) obtained so far by the TIE* algorithm and ordered by the time of discovery from latest to earliest;
- subsets $\mathbf{G}_1,...,\mathbf{G}_n$ that were removed from **V** (in line 5 of TIE*) to obtain the above Markov boundaries ($\mathbf{G}_1=\varnothing$);
- subsets $\mathbf{G}_1^*,...,\mathbf{G}_m^*$ that were removed from **V** (in line 5 of TIE*) and did not lead to Markov boundaries in the original distribution (in line 6 of TIE*).

Output: $\mathbf{G} \subseteq \bigcup_{i=1}^{n} \mathbf{M}_i$

Generate a subset of variables $\mathbf{G} \subseteq \mathbf{M}_i \cup \mathbf{G}_i$ (i = 1, .., n) with the smallest number of variables and the smallest lexicographical order (for strategy *IncLex*) or the smallest association with *T* (for strategy *IncMinAssoc*) or the largest association with *T* (for strategy *IncMaxAssoc*) such that:

$$\mathbf{G} \supset \mathbf{G}_i,\ \mathbf{G} \not\supseteq \mathbf{G}_j^*, \text{ and } \mathbf{G} \neq \mathbf{G}_k \text{ for } j=1, \ldots, m \text{ and } k=1, \ldots, n.$$

Figure 13: Strategies *IncLex*, *IncMinAssoc*, and *IncMaxAssoc* to generate subsets of variables that have to be removed from **V** to identify new Markov boundaries of *T*. "*Inc*" in the name of the strategy stands for incremental generation of subsets; "*Lex*" stands for lexicographical order; "*MinAssoc*" stands for minimal association with *T*; and "*MaxAssoc*" stands for maximal association with *T*.

Proof: Consider that the strategy in Figure 13 generated a subset $\mathbf{G}' \subseteq \bigcup_{i=1}^{n} \mathbf{M}_i$ leading to identification of a Markov boundary $\mathbf{M}_{new}$ in the embedded distribution in line 5 of the TIE* algorithm. Assume that $\mathbf{M}_{new}$ is not a Markov blanket in the original distribution. Thus, it is not a Markov boundary in the original distribution. Since removal of **G'** does not lead to a Markov boundary in the original distribution, the strategy dictates not to generate supersets of **G'**.

Assume that there is a set **W** that is a Markov boundary of T in the original distribution and it is not output by TIE* because **G"**: **G"** $\supset$ **G'** was not generated.

- Since **W** is a Markov blanket of T in the original distribution and $\mathbf{M}_{new}$ is not, Theorem 1 implies that performance of a learning algorithm $\mathbb{L}$ that can approximate any probability distribution for prediction of T measured by metric $\mathbb{M}$ that is maximized only when P(T | **V** \ {T}) is estimated accurately is larger for **W** than for $\mathbf{M}_{new}$.
- Since **W** satisfies $T \perp (\mathbf{V} \setminus \mathbf{W} \setminus \{T\}) \mid \mathbf{W}$ by the definition of Markov blanket, decomposition property implies that $T \perp (\mathbf{V} \setminus \mathbf{W} \setminus \mathbf{G}' \setminus \{T\}) \mid \mathbf{W}$, i.e. **W** similarly to $\mathbf{M}_{new}$ is a Markov blanket of T in the embedded distribution after removal of **G'**. Therefore by Theorem 1, performance of a learning algorithm $\mathbb{L}$ that can approximate any probability distribution for prediction of T measured by metric $\mathbb{M}$ that is maximized only when P(T | **V** \ {T}) is estimated accurately should be the same for **W** and $\mathbf{M}_{new}$.

  The above two points are contradictory, thus **W** does not exist. (Q.E.D.)

On the choice of admissible input components for TIE*

The above subsections presented several examples of admissible input components for the TIE* algorithm that satisfy rules given in Figure 6. I would like to reiterate that the input components are selected to be suitable for the distribution in hand and should satisfy admissibility rules for correctness of the TIE* algorithm.

If the underlying distribution satisfies the local composition property, then IAMB can be used as an admissible Markov boundary learner (input component $\mathbb{X}$). If the distribution satisfies a relaxed version of the faithfulness and other assumptions of Theorem 9, then HITON-PC can be used. Many other Markov boundary learners can also be proven admissible given specific distributional assumptions. The next chapter presents empirical results of using a Markov boundary learner that does not require faithfulness assumption.

The choice of the strategy to generate subsets of variables that have to be removed to identify new Markov boundaries (input component $\mathbb{Y}$) is also dependent on the distribution. In general, the admissible strategies outlined in Figure 13 should be suitable for all distributions, but they are not necessarily the most computationally efficient ones. For example, in some distributions it may be sufficient to consider only subsets of the Markov boundary **M** that is discovered in line 1 of the TIE* algorithm. For other distributions, it may be sufficient to consider subsets of variables limited up to certain size. Other distributions may also require for additional computational efficiency removal of subsets that are not limited to the Markov boundary members.

Finally, the criterion for verification of Markov boundaries (input component $\mathbb{Z}$) has also to be selected for the distribution in hand. I outlined and proved admissibility for two such criteria: one uses conditional independence tests (Figure 11) and the other applies a learning algorithm and assesses predictivity using a formal statistical test (Figure 12). The choice between these two criteria can be dictated by available sample size, size of the Markov boundaries, difficulty of the learning problems, and so on. Other domains may also require use of different verification criteria.

On the computational complexity of TIE*

The computational complexity of the TIE* algorithm depends both on the specific instantiations of the input components ($\mathbb{X}$, $\mathbb{Y}$, $\mathbb{Z}$) and on the underlying distribution. One of the most computationally expensive steps of TIE* is learning Markov boundaries (i.e., using input component $\mathbb{X}$). In the above subsections, I have described two Markov boundary learning algorithms, IAMB (Figure 8) and HITON-PC (Figure 9). The computational complexity of these algorithms is usually measured by the number of conditional independence tests. The average-case complexity of IAMB is $O(|\mathbf{V}||\mathbf{M}|)$, and the complexity of HITON-PC is $O(|\mathbf{V}|2^{|\mathbf{M}|})$, where

|**V**| is the number of variables in the dataset and |**M**| is the number of variables in the tentative Markov boundary which is typically of the same order as the number of variables in the true Markov boundary. Assuming that there are t true Markov boundaries in the distribution, that each of them has |**M**| variables, and that it takes TIE* O(t) runs of the Markov boundary learner to find these Markov boundaries, the overall computational complexity of TIE* is O(t|**V**||**M**|) and O(t|**V**|$2^{|\mathbf{M}|}$) conditional independence tests when using IAMB and HITON-PC, respectively. The above estimates do not take into account computational expenses incurred by using input components $\mathbb{Y}$ and $\mathbb{Z}$ in the TIE* algorithm. However, in practical applications >95-99% of CPU time is spent on learning Markov boundaries which justifies use of the above estimates.

CHAPTER V

EMPIRICAL EVALUATION IN ARTIFICIAL SIMULATED DATA

Before applying TIE* to real data, I test its behavior in artificially simulated datasets where all Markov boundaries (and thus all maximally predictive and non-redundant signatures) are known. This allows to test whether the algorithm behaves according to theoretical expectations and study its empirical properties. This also provides clues about the behavior of TIE* and the baseline comparison algorithms in the experiments with real human microarray data.

Many of the reported experiments involve the following four performance metrics:

- γ = total number of Markov boundaries output by the algorithm (not necessarily correctly);
- ω = number of Markov boundaries that were correctly discovered (relative to the gold standard) with no false negative variables but with possible false positive (redundant) variables;
- ϕ = average number of false positive variables in discovered Markov boundaries that were used for computation of ω;
- δ = penalized proportion of discovered Markov boundaries that is computed as follows: For every true Markov boundary Θ_i ($i = 1, \ldots, N$), find a Markov boundary output by the algorithm that maximizes the product of sensitivity and specificity for identification of this true Markov boundary: $\alpha_i = \max_k \left(\frac{|\mathbf{M}_k \cap \Theta_i|}{|\Theta_i|} \left(1 - \frac{|\mathbf{M}_k \setminus \Theta_i|}{|\mathbf{V} \setminus (\Theta_i \cup \{T\})|} \right) \right)$, where $\mathbf{M}_k$ is a Markov boundary output by the algorithm. Once such Markov boundary is

identified, it is not considered again for computation of α_i for the next true Markov boundary. Finally, δ is defined as $\frac{1}{N}\sum_{i=1}^{N}\alpha_i$.

Experiments with discrete networks *TIED1* and *TIED2*

There are two goals of the experiments reported in this section: (i) to analyze behavior of the TIE* algorithm as a function of sample size using data generated from a discrete network (experiment 1) and (ii) to compare TIE* to state-of-the-art algorithms and examine sensitivity of the tested methods to high dimensionality (experiment 2).

Two discrete networks denoted as *TIED1* and *TIED2* were constructed with 30 and 1,000 variables, respectively. Both networks have the same 72 Markov boundaries. The details about network structure and parameterization are provided in Appendix D.

The following instantiation of the TIE* algorithm was used in experiments. It can be described by a tuple of input components (𝕏, 𝕐, ℤ):

- 𝕏 (Markov boundary algorithm) = HITON-PC that uses G^2 test with α = 0.05 (Figure 9);
- 𝕐 (strategy to generate subsets of variables that have to be removed to identify new Markov boundaries of *T*) = *IncLex* (Figure 13);
- ℤ (criterion to verify Markov boundaries) = *Independence* that uses G^2 test with α = 0.05 (Figure 11).

In experiment 2, eight state-of-the-art algorithms were used to compare to TIE* as described in Appendix G.

Experiment 1: This experiment involved running TIE* to discover all Markov boundaries of *T* in training datasets of different sample sizes generated from the *TIED1* network. Ten samples of each size were used to reduce variability in the reported results.

Metric	Sample size					
	200	300	500	1000	2000	5000
γ	43.2	21.6	101.2	72	72	72
ω	21.6	16.2	72	72	72	72
ϕ	4.4	3	0.1	0	0	0
δ	0.391	0.208	0.996	1	1	1

Table 1: Results of experiment 1 with artificial dataset from *TIED1* network. Performance metrics are averaged over 10 samples of each size. 72 true Markov boundaries (i.e., maximally predictive and non-redundant signatures) exist in this distribution.

As can be seen in Table 1, the algorithm identifies all 72 true Markov boundaries with very few false positive variables and with ~30 false positive Markov boundaries when the sample size is 500; and when the sample size is $\geq$ 1,000 the algorithm outputs all 72 true Markov boundaries exactly. In fact, even when the sample size is 750, the algorithm does not make any errors in its output (data not shown).

Experiment 2: This experiment involved running TIE* and baseline comparison algorithms to discover all Markov boundaries of *T* (i.e., maximally predictive and non-redundant signatures) in two training datasets with 750 samples each, generated from *TIED1* and *TIED2* networks. Once variables that participate in the signatures were identified, a one-versus-rest multicategory linear SVM classifier (Schölkopf et al., 1999; Vapnik, 1998) was trained in the training dataset and tested in the non-overlapping 3,000 sample independent validation dataset. The predictive performance was measured by the weighted accuracy metric (Guyon et al., 2006).

Tables 2 and 3 present results of the experiment. The following are observed: (i) TIE* perfectly identifies all 72 true Markov boundaries (maximally predictive and non-redundant signatures) in the datasets with either 30 or 1,000 variables; (ii) Iterative Removal identifies only 1 signature because all other signatures have common variables and thus cannot be detected by this method; (iii) KIAMB fails to identify any true signature due to its sample inefficiency, and because of the same reason its signatures have poor predictivity; (iv) resampling-based methods either miss many true signatures and/or output many redundant variables in the signatures.

Method	Total number of output signatures (γ)	Number of variables in an average output signature	Number of Markov boundaries (i.e., true signatures)		Average number of redundant variables in identified true signatures (ϕ)	Average predictive performance in validation data	CPU time in minutes
			identified exactly	identified with redundant variables (ω)			
TIE*	72	5.00	72	72	0.00	0.951	0.39
Iterative Removal	3	4.67	0	1	1.00	0.946	0.01
KIAMB1	5000	2.83	0	0	N/A	0.776	11.55
KIAMB2	5000	2.82	0	0	N/A	0.772	11.69
KIAMB3	5000	2.81	0	0	N/A	0.774	11.62
Resampling+Univariate1	5000	17.87	0	72	12.00	0.949	84.56
Resampling+Univariate2	5000	7.54	0	25	12.12	0.924	85.50
Resampling+RFE1	5000	14.25	0	72	5.01	0.954	78.71
Resampling+RFE2	5000	5.80	1	44	4.25	0.939	79.26

Table 2: Results of experiment 2 with artificial dataset from *TIED1* network (with 30 variables). 72 true Markov boundaries (i.e., maximally predictive and non-redundant signatures) exist in this distribution. The predictive performance is measured by the weighted accuracy metric. The optimal Bayes classification performance is 0.9663 (weighted accuracy).

Method	Total number of output signatures (γ)	Number of variables in an average output signature	Number of Markov boundaries (i.e., true signatures)		Average number of redundant variables in identified true signatures (ϕ)	Average predictive performance in validation data	CPU time in minutes
			identified exactly	identified with redundant variables (ω)			
TIE*	72	5.00	72	72	0.00	0.957	0.46
Iterative Removal	3	5.67	0	1	2.00	0.959	0.04
KIAMB1	5000	2.82	0	0	N/A	0.798	285.42
KIAMB2	5000	2.81	0	0	N/A	0.796	285.45
KIAMB3	5000	2.80	0	0	N/A	0.796	285.48
Resampling+Univariate1	5000	11.10	0	72	12.29	0.942	5999.64
Resampling+Univariate2	5000	5.58	0	0	N/A	0.934	6000.41
Resampling+RFE1	5000	8.70	0	72	6.38	0.952	6235.28
Resampling+RFE2	5000	4.24	0	29	5.76	0.947	6235.93

Table 3: Results of experiment 2 with artificial dataset from *TIED2* network (with 1,000 variables). 72 true Markov boundaries (i.e., maximally predictive and non-redundant signatures) exist in this distribution. The predictive performance is measured by the weighted accuracy metric.

Experiments with linear continuous network *LIND*

There are two goals of the experiments reported in this section: (i) to analyze behavior of the TIE* algorithm as a function of sample size using data generated from a continuous network and (ii) to compare criteria *Independence* (Figure 11) and *Predictivity* (Figure 12) for verification of Markov boundaries in the TIE* algorithm.

A continuous network denoted as *LIND* was constructed with 41 variables. There are 12 Markov boundaries in the network. The details about network structure and parameterization are provided in Appendix E.

The following two instantiations of the TIE* algorithm were used in experiments. They can be described by tuples of input components ($\mathbb{X}$, $\mathbb{Y}$, $\mathbb{Z}^1$) and ($\mathbb{X}$, $\mathbb{Y}$, $\mathbb{Z}^2$), respectively:

- $\mathbb{X}$ (Markov boundary algorithm) = HITON-PC (Figure 9) that uses Fisher's Z test with $\alpha = 0.05$;
- $\mathbb{Y}$ (strategy to generate subsets of variables that have to be removed to identify new Markov boundaries of T) = *IncLex* (Figure 13);
- $\mathbb{Z}^1$ (criterion to verify Markov boundaries) = *Independence* (Figure 11) that uses G^2 test with $\alpha = 0.05$;
- $\mathbb{Z}^2$ (criterion to verify Markov boundaries) = *Predictivity* (Figure 12) that uses:
    - $\mathbb{L}$ = linear SVM classifier (Vapnik, 1998);
    - $\mathbb{M}$ = area under ROC curve (AUC) performance metric (Fawcett, 2003);
    - $\mathbb{E}$ = holdout validation performance estimator;
    - $\mathbb{C}$ = nonparametric method to compare estimates of AUC with $\alpha = \{0.1, 0.05, 0.01, 0.005, 0.001\}$ (DeLong et al., 1988).

γ

Sample size	Criterion *Independence* (Z^1)	Criterion *Predictivity* (Z^2) with α for Delong's test =				
		0.1	0.05	0.01	0.005	0.001
200	10.4	26.7	36.1	56.4	64.6	77.4
300	11.7	24.3	27.9	39.4	47.3	63.6
500	10.8	12.3	15.6	20.6	24.8	40.7
1000	11.1	10.2	12	12	12	16.3

ω

Sample size	Criterion *Independence* (Z^1)	Criterion *Predictivity* (Z^2) with α for Delong's test =				
		0.1	0.05	0.01	0.005	0.001
200	10.4	10.8	11.4	11.4	11.4	11.4
300	11.4	11.4	12	12	12	12
500	10.8	10.5	12	12	12	12
1000	11.1	10.2	11.4	11.4	11.4	12

ϕ

Sample size	Criterion *Independence* (Z^1)	Criterion *Predictivity* (Z^2) with α for Delong's test =				
		0.1	0.05	0.01	0.005	0.001
200	0.09	0.05	0.05	0.05	0.05	0.05
300	0.33	0.25	0.25	0.25	0.25	0.25
500	0.15	0.1	0.1	0.1	0.1	0.1
1000	0	0	0	0	0	0

δ

Sample size	Criterion *Independence* (Z^1)	Criterion *Predictivity* (Z^2) with α for Delong's test =				
		0.1	0.05	0.01	0.005	0.001
200	0.81	0.93	0.99	0.99	0.99	0.99
300	0.93	0.98	1	1	1	1
500	0.85	0.84	1	1	1	1
1000	0.89	0.81	0.94	0.94	0.94	1

Table 4: Results of experiments with artificial dataset from *LIND* network (with 41 variables). Performance metrics are averaged over 10 samples of each size. 12 true Markov boundaries (i.e., maximally predictive and non-redundant signatures) exist in this distribution.

Both instantiations of the TIE* algorithm were run to discover all Markov boundaries of *T* in training datasets of different sample sizes generated from the *LIND* network. Ten samples of each size were used to reduce variability in the reported results.

Table 4 shows results of the experiments. The following are observed: (i) as sample size increases, the performance of both instantiations of TIE* (as measured by ω and ϕ) generally improves and the algorithms discover up to 11 or 12 (all) true Markov boundaries; (ii) the α-level in the criterion *Predictivity* significantly affects the number of Markov boundaries output by the TIE* algorithm: the smaller is α, the more Markov boundaries are output; (iii) TIE* with the criterion *Predictivity* typically leads to a larger number of output Markov boundaries than with the criterion *Independence*; (iv) TIE* with the criterion *Predictivity* in most cases and on average leads to superior performance (as measured by ω, ϕ, and δ) compared to the criterion *Independence*.

The latter finding suggests use of TIE* with the criterion *Predictivity* in experiments with microarray gene expression data, especially given that the criterion *Independence* may be based on unreliable statistical tests when the sample size is small.

Experiments with discrete network *XORD*

The experiments reported in this section seek to evaluate TIE* when popular Markov boundary learners such as IAMB (Figure 8) and HITON-PC (Figure 9) are not applicable due to violations of their fundamental assumptions. Specifically, the behavior of TIE* is examined when the local composition property with respect to response variable *T* (and thus faithfulness) is violated. The generative nature of the TIE* algorithm allows to select and use a Markov boundary learner suitable for the distribution in hand.

A discrete network denoted as *XORD* was constructed with 41 variables. There are 25 true Markov boundaries in the network. The details about network structure and parameterization are provided in Appendix F.

The following instantiation of the TIE* algorithm was used in experiments. It can be described by a tuple of input components ($\mathbb{X}$, $\mathbb{Y}$, $\mathbb{Z}$):

- $\mathbb{X}$ (Markov boundary algorithm) = heuristic algorithm SVM-FSMB (Tsamardinos and Brown, 2008) that uses HITON-MB (Aliferis et al., 2008a) with G^2 test and $\alpha = 0.05$ in the SVM feature space;
- $\mathbb{Y}$ (strategy to generate subsets of variables that have to be removed to identify new Markov boundaries of T) = *IncLex* (Figure 13);
- $\mathbb{Z}$ (criterion to verify Markov boundaries) = *Predictivity* (Figure 12) that uses:
    - $\mathbb{L}$ = polynomial SVM classifier of degree 3 (Vapnik, 1998);
    - $\mathbb{M}$ = area under ROC curve (AUC) performance metric (Fawcett, 2003);
    - $\mathbb{E}$ = holdout validation performance estimator;
    - $\mathbb{C}$ = nonparametric method to compare estimates of AUC with $\alpha = 0.1$[4] (DeLong et al., 1988).

TIE* was run to discover all Markov boundaries of T in training datasets of different sample sizes generated from the *XORD* network. Ten samples of each size were used to reduce variability in the reported results.

Table 5 reports results of the experiments. The following are observed: (i) TIE* can discover all 25 true Markov boundaries when the sample is ≥ 2,000; (ii) there is ~1 false positive variable in each discovered Markov boundary for large sample sizes; (iii) TIE* discovers only

[4] I experimented with several α-levels {0.1, 0.05, 0.01, 0.005, 0.001} for the Delong's test. The results appear to be insensitive to the choice of α-level, and thus I report results only for a single α-level. The reason for this insensitivity is dramatic difference of observed classification AUC's: e.g. when a true Markov blanket is discovered its AUC ≈ 1, otherwise AUC ≈ 0.5.

one Markov boundary with the smallest number of variables $\{X_9, X_{10}, X_{11}\}$ when sample size is 300 due to inability of SVM-FSMB to output more Markov boundaries for that sample size.

Metric	Sample size				
	300	500	1000	2000	5000
γ	1	1.2	9.3	58.7	43
ω	1	1.2	5.2	25	25
ϕ	0	0.13	0.33	1.11	0.58
δ	0.04	0.05	0.23	1	1

Table 5: Results of experiments with artificial dataset from *XORD* network. Performance metrics are averaged over 10 samples of each size. 25 true Markov boundaries (i.e., maximally predictive and non-redundant signatures) exist in this distribution.

CHAPTER VI

EMPIRICAL EVALUATION IN RESIMULATED MICROARRAY GENE EXPRESSION DATA

The experiments in the previous chapter demonstrated excellent empirical properties of TIE* in several artificial simulated datasets. However, one can argue that these distributions may be different from real microarray gene expression data. Therefore, I extend evaluation of the TIE* algorithm to resimulated microarray gene expression data that by design closely resembles real microarray data. The knowledge of a generative model for this dataset allows to generate arbitrary large samples from the distribution and study the behavior of TIE* as a function of sample size. However, unlike the experiments with the artificial simulated datasets, all maximally predictive and non-redundant signatures are not known.

There are five goals of the experiments reported in this chapter: (i) to examine whether the signature multiplicity phenomenon vanishes as the sample size grows (experiment 1); (ii) to assess stability of TIE* to the initial signature **M** that is obtained in line 1 of the algorithm (experiment 2); (iii) to experiment with several wrapping strategies as an additional post-processing step for the TIE* signatures in order to increase their number and maximize their parsimony (experiment 3); (iv) to compare TIE* with baseline algorithms (also experiment 3); and (v) to examine the relative contribution of other signatures to the ones output by TIE* (also experiment 3).

A resimulated gene expression network with 1,000 variables (999 genes and a phenotypic response variable) was reverse-engineered and the data was generated as described in Appendix H.

The following instantiation of the TIE* algorithm was used in experiments. It can be described by a tuple of input components ($\mathbb{X}$, $\mathbb{Y}$, $\mathbb{Z}$):

- $\mathbb{X}$ (Markov boundary algorithm) = HITON-PC (Figure 9) that uses Fisher's Z test with $\alpha = 0.05$;
- $\mathbb{Y}$ (strategy to generate subsets of variables that have to be removed to identify new Markov boundaries of *T*) = *IncLex* (Figure 13) with the maximum size of subset **G** limited to 5 genes (in experiments 1 and 3) or 4 genes (in experiment 2)[5].
- $\mathbb{Z}$ (criterion to verify Markov boundaries) = *Predictivity* (Figure 12) that uses:
    - $\mathbb{L}$ = linear SVM classifier (Vapnik, 1998);
    - $\mathbb{M}$ = area under ROC curve (AUC) performance metric (Fawcett, 2003);
    - $\mathbb{E}$ = holdout validation performance estimator;
    - $\mathbb{C}$ = nonparametric method to compare estimates of AUC with $\alpha = 0.1$ (DeLong et al., 1988).

In experiment 3, eight state-of-the-art algorithms were used to compare to TIE* as described in Appendix G.

Experiment 1: TIE* was applied to resimulated gene expression data with sample sizes: 300, 450, …, 1,500, 2,250, 3,000, … 30,000. The number of unique signatures and the number of unique non-reducible[6] signatures discovered by the algorithm for each sample size is shown in Figure 14. The discovered signatures were maximally predictive of *T* as confirmed by holdout validation. As sample size increases, the number of output signatures drops but then remains constant in the range 160-644 (or 53-279 for non-reducible signatures) for datasets with ≥ 4,500 samples. *This shows the existence of at least two sources of multiplicit*y: one is small sample size

[5] Note that this can lead to recovery of only a fraction of all maximally predictive and non-redundant signatures while making the experiments computationally feasible.

[6] A signature is called *non-reducible* if it is not properly included in any other output signature (i.e., it is a proxy of having no redundant genes). For example, if a method outputs 3 signatures with the following genes: {*A*, *B*, *C*}, {*A*, *B*, *X*}, and {*A*, *B*}, only signature {*A*, *B*} is non-reducible.

and the other is multiplicity intrinsic to gene-gene and gene-phenotype relations. As sample size grows, the first source vanishes and only the second one remains. Since the resimulated data distribution closely resembles the real-life distribution (see Appendix H), this experiment supports the hypothesized existence of multiple signatures in very large samples (>10,000) contrary to the theoretical model of (Ein-Dor et al., 2006).

It is also worthwhile to note that the resimulated network from this experiment was obtained from real microarray data using methods that rely on the faithfulness assumption and therefore induce only a single local neighborhood for each variable. Thus, the true multiplicity of signatures may be much larger than the results presented in Figure 14.

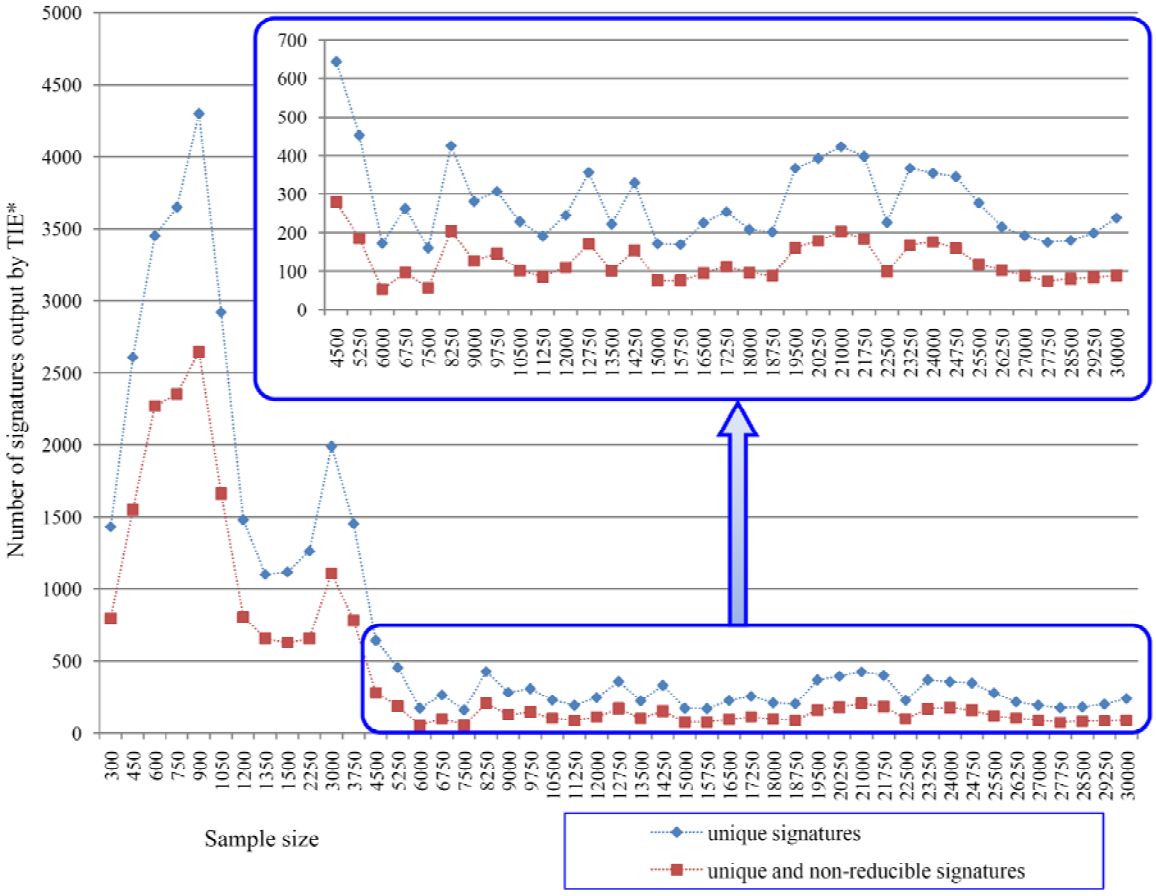


Figure 14: Number of maximally predictive signatures output by TIE* as sample size grows. The inner figure is a magnified region of the main figure.

Experiment 2: In this experiment TIE* was first applied to a sample of the size 1,000 to discover all maximally predictive and non-redundant signatures. The majority of discovered signatures contained 9-13 genes, and the initial signature **M** that is obtained in line 1 of the algorithm contained 12 genes. For each of the five most common sizes of signatures {9, 10, 11, 12, 13} five signatures ("seeds") were randomly selected from the output of TIE*. These seed signatures were then used in TIE* instead of the initial signature **M** (i.e., TIE* algorithm was rerun for each seed signature starting from line 2). The above experiment was repeated on five samples of the size 1,000 to minimize variability in the reported results.

In Table 6 I assess stability of TIE* to the choice of initial signature **M** by computing a proportion of signatures output by TIE* run with the seed signature that belong to the output of TIE* run with the initial signature **M** (I denote this proportion by λ). As can be seen, TIE* exhibits exceptional stability, therefore the choice of seed signature does not affect the output of the algorithm. This is a very important finding because Markov boundary algorithms such as HITON-PC and IAMB guarantee to output a maximally predictive and non-redundant signature which can potentially be any of such multiple signatures that exist in the distribution.

Metric λ	**Size of seed signature**				
	9	10	11	12	13
average	99.32%	99.54%	99.51%	99.44%	99.52%
min	97.69%	98.60%	98.33%	98.20%	98.21%
max	100.00%	100.00%	100.00%	100.00%	100.00%

Table 6: Analysis of stability of TIE* to the choice of initial signature **M**. Metric λ denotes a proportion of signatures output by TIE* run with the seed signature that belong to the output of TIE* run with the initial signature **M**. The reported values of metric λ are first averaged over 5 seed signatures of each size and then either averaged or minimized or maximized over 5 samples as shown in the table.

Experiment 3: This experiment first involved running TIE* and baseline comparison algorithms to discover all maximally predictive and non-redundant signatures in five samples of

sizes {200, 300, 500, 1,000} each. Once TIE* has output signatures, they were post-processed with one of the following three wrapping strategies (Kohavi and John, 1997):

- *Wrapping1*: For each output TIE* signature, sort its genes by univariate association with response variable and perform backward wrapping to create a new signature. Output unique signatures;
- *Wrapping2*: For each output TIE* signature, sort its genes randomly and perform backward wrapping to create a new signature. Output unique signatures;
- *Wrapping3*: For each output TIE* signature, sort its genes randomly 50 times and perform backward wrapping for each random ordering to create new signatures. Output unique signatures.

The above three wrapping procedures give rise to the methods denoted as TIE* + Wrapping1, TIE* + Wrapping2, and TIE* + Wrapping3, respectively. Once genes that constitute the signatures were identified, a linear SVM classifier (Vapnik, 1998) was trained and tested by holdout validation. The predictive performance was measured by area under ROC curve (AUC) metric (Fawcett, 2003).

The results for the total number of unique signatures, number of genes in an average signature, and average holdout validation predictive performance of signatures output by each algorithm are provided in Tables 7, 8, and 9, respectively. As can be seen: (i) all wrapping strategies applied to TIE* result in more parsimonious signatures at the expense of a slight decrease of predictive performance that is not statistically significant in each dataset; (ii) TIE* + Wrapping3 results in more signatures compared with TIE* when the sample size is $\geq$ 500; and (iii) all other methods typically output signatures that are either less parsimonious and/or have inferior predictive performance.

The remainder of this experiment is devoted to assessing the value of other signatures relative to the ones output by TIE* or TIE*+Wrapping. All unique signatures output by tested methods were considered. The signatures that have statistically maximal predictive performance

and that are also non-reducible relative to all output signatures were denoted as “positives”. All other signatures were denoted as “negatives”. This allowed computation of sensitivity and specificity for each algorithm.

The results for TIE* + Wrapping1, TIE* + Wrapping2, TIE* + Wrapping3 for sample size 1,000 are reported in Tables 10, 11, 12 respectively. As can be seen: (i) all TIE* algorithms are much closer to the point with sensitivity = 1 and specificity = 1 than other non-TIE* methods; (ii) TIE* + Wrapping1 maximizes specificity (=0.99) while having sensitivity (=0.77) superior to other methods; (iii) TIE* + Wrapping3 maximizes sensitivity (=0.99) and has very good specificity (=0.64); and (iv) TIE* + Wrapping2 simultaneously achieves excellent sensitivity (=0.96) and specificity (=0.94).

These findings suggest that signatures output by tested non-TIE* methods are either redundant or have inferior predictivity compared to signatures output by TIE* techniques. In general, use of wrapping algorithms on top of TIE* signatures may not be needed if (i) the sample size is large enough for conditional independence tests to be reliable (otherwise a Markov boundary learner may include redundant variables in its output as demonstrated in Appendix J) and (ii) the predictive performance metric is maximized only when $P(T \mid \mathbf{V} \setminus \{T\})$ is estimated accurately.

Algorithm	Sample size			
	200	300	500	1000
TIE*	796	1180.8	2957.6	3926.2
TIE* + Wrapping1	22.4	61.6	129.8	300.8
TIE* + Wrapping2	102.6	204.2	763.6	1365.8
TIE* + Wrapping3	326.2	614.6	3090.2	7148
Iterative Removal	2.2	2.2	1	1
Resampling + RFE1	3260.8	3503.6	3468	4505.6
Resampling + RFE2	663.4	732.4	537.8	1207.2
Resampling + UAF1	2313.2	2444.6	2361.6	3136.4
Resampling + UAF2	328	322	201.2	377.6
KIAMB1	1151.6	1107.4	655.8	420.6
KIAMB2	182.8	224.6	208.6	139
KIAMB3	14.2	27.2	34	28.6

Table 7: Total number of unique signatures output by algorithms (averaged over 5 samples).

Algorithm	Sample size			
	200	300	500	1000
TIE*	6.92	7.78	9.99	10.85
TIE* + Wrapping1	1.42	1.53	2.68	4.09
TIE* + Wrapping2	1.65	1.83	3.21	4.47
TIE* + Wrapping3	3.02	3.54	4.89	6.13
Iterative Removal	7.33	7.73	11.6	12.2
Resampling + RFE1	15.4	15.66	16.47	36.32
Resampling + RFE2	1.86	2.26	2.38	4.48
Resampling + UAF1	19.35	20.09	24.67	55.19
Resampling + UAF2	1.66	2.15	2.24	4.32
KIAMB1	21.21	24.18	25.18	31.22
KIAMB2	12.57	15.16	18.94	24.62
KIAMB3	7.2	9.57	11.89	18.01

Table 8: Number of genes in an average signature output by algorithms (averaged over 5 samples).

Algorithm	Sample size			
	200	300	500	1000
TIE*	0.985	0.988	0.997	0.999
TIE* + Wrapping1	0.929	0.930	0.968	0.988
TIE* + Wrapping2	0.920	0.920	0.967	0.987
TIE* + Wrapping3	0.927	0.927	0.969	0.988
Iterative Removal	0.986	0.971	1.000	1.000
Resampling + RFE1	0.975	0.984	0.992	0.999
Resampling + RFE2	0.908	0.919	0.942	0.984
Resampling + UAF1	0.964	0.977	0.991	0.999
Resampling + UAF2	0.908	0.930	0.942	0.985
KIAMB1	0.987	0.993	0.998	0.999
KIAMB2	0.978	0.994	0.998	0.999
KIAMB3	0.982	0.994	0.997	0.999

Table 9: Average holdout validation predictive performance (AUC) of signatures output by algorithms (averaged over 5 samples).

Algorithm	Sensitivity	Specificity	Distance
TIE* + Wrapping1	0.77	0.99	0.23
Iterative Removal	0.00	1.00	1.00
Resampling + RFE1	0.02	0.53	1.09
Resampling + RFE2	0.26	0.88	0.75
Resampling + UAF1	0.01	0.68	1.04
Resampling + UAF2	0.05	0.96	0.95
KIAMB1	0.00	0.96	1.00
KIAMB2	0.00	0.99	1.00
KIAMB3	0.00	1.00	1.00

Table 10: Comparison of TIE* + Wrapping1 with all other non-TIE* methods in terms of sensitivity, specificity, and Euclidian distance (from point with sensitivity = 1 and specificity = 1 in the ROC space) for detection of the set of maximally predictive and non-redundant signatures. The reported results are averaged over 5 samples of size 1,000.

Algorithm	Sensitivity	Specificity	Distance
TIE* + Wrapping2	0.96	0.94	0.07
Iterative Removal	0.00	1.00	1.00
Resampling + RFE1	0.01	0.55	1.09
Resampling + RFE2	0.07	0.88	0.94
Resampling + UAF1	0.00	0.69	1.04
Resampling + UAF2	0.01	0.96	0.99
KIAMB1	0.00	0.96	1.00
KIAMB2	0.00	0.99	1.00
KIAMB3	0.00	1.00	1.00

Table 11: Comparison of TIE* + Wrapping2 with all other non-TIE* methods in terms of sensitivity, specificity, and Euclidian distance (from point with sensitivity = 1 and specificity = 1 in the ROC space) for detection of the set of maximally predictive and non-redundant signatures. The reported results are averaged over 5 samples of size 1,000.

Algorithm	Sensitivity	Specificity	Distance
TIE* + Wrapping3	0.99	0.64	0.36
Iterative Removal	0.00	1.00	1.00
Resampling + RFE1	0.00	0.69	1.04
Resampling + RFE2	0.03	0.92	0.98
Resampling + UAF1	0.00	0.79	1.02
Resampling + UAF2	0.00	0.98	1.00
KIAMB1	0.00	0.97	1.00
KIAMB2	0.00	0.99	1.00
KIAMB3	0.00	1.00	1.00

Table 12: Comparison of TIE* + Wrapping3 with all other non-TIE* methods in terms of sensitivity, specificity, and Euclidian distance (from point with sensitivity = 1 and specificity = 1 in the ROC space) for detection of the set of maximally predictive and non-redundant signatures. The reported results are averaged over 5 samples of size 1,000.

CHAPTER VII

EMPIRICAL EVALUATION IN REAL HUMAN MICROARRAY GENE EXPRESSION DATA

The results of experiments in several simulated artificial datasets and the resimulated microarray dataset described in the previous two chapters showcase excellent empirical performance of the TIE* algorithm and its advantages over state-of-the-art methods. In the current chapter I extend the evaluation of multiple signature extraction methods to real human microarray gene expression data where maximally predictive and non-redundant signatures are not known *a priori*, and the data generative functions are not known as well. The major emphasis of this chapter are independent-dataset experiments than involve two microarray datasets either from different laboratories or different platforms; one is used for discovery of signatures and another is used for their validation. Even though evaluation of multiple signature extraction methods using independent-dataset design can be considered convincing by many practitioners, it is challenging due to potential differences in sample populations between the two datasets. That is why I also included experiments with relatively large sample size microarray datasets that can be used both for discovery and validation of signatures.

Independent-dataset experiments

The primary goal of experiments reported in this section is to compare TIE* and baseline algorithms for extraction of multiple signatures in terms of maximal predictivity of induced signatures and reproducibility in independent data. Operationally, I define maximal predictivity (classification performance) for each dataset as follows: I apply all tested methods for extraction of multiple signatures to some dataset; then for each method I compute average predictivity of the phenotype (over all identified signatures by this method) measured by area under ROC curve;

finally I compute the maximum value of the above average predictivity estimates and refer to it as *maximal predictivity*.

In these experiments, I adopted an independent-dataset design where one microarray dataset ("*discovery dataset*") was used for identification of signatures and estimation of their predictivity by holdout validation, and another independent dataset ("*validation dataset*") originating either from a *different laboratory* or from a *different microarray platform* was used for validation of predictivity of the signatures. No overlap of samples between discovery and validation dataset analyses occurs in this design. The criteria for dataset admissibility and protocol for quality assurance and processing of pairs of datasets are described in Appendix I. In total, 6 pairs of gene expression microarray datasets covering both human cancer diagnosis and clinical outcome prediction were used (listed in Table 13).

The following instantiation of the TIE* algorithm was used in experiments. It can be described by a tuple of input components ($\mathbb{X}$, $\mathbb{Y}$, $\mathbb{Z}$):

- $\mathbb{X}$ (Markov boundary algorithm) = HITON-PC (Figure 9) that uses Fisher's Z test with $\alpha = 0.05$;
- $\mathbb{Y}$ (strategy to generate subsets of variables that have to be removed to identify new Markov boundaries of T) = *IncLex* (Figure 13) with the maximum size of subset **G** limited to 5 genes.
- $\mathbb{Z}$ (criterion to verify Markov boundaries) = *Predictivity* (Figure 12) that uses:
    - $\mathbb{L}$ = linear SVM classifier (Vapnik, 1998);
    - $\mathbb{M}$ = area under ROC curve (AUC) performance metric (Fawcett, 2003);
    - $\mathbb{E}$ = holdout validation performance estimator;
    - $\mathbb{C}$ = nonparametric method to compare estimates of AUC with $\alpha = 0.1$ (DeLong et al., 1988).

Task	*Discovery dataset*					*Validation dataset*					Number of common genes
	Reference	Sample size	Samples per class	Number of genes	Microarray platform	Reference	Sample size	Samples per class	Number of genes	Microarray platform	
Lung Cancer Diagnosis: lung tumors vs. normals (non-tumor lung samples)	(Bhattacharjee et al., 2001)	203	lung tumors (**186**) normals (**17**)	12600	Affymetrix U95A	(Beer et al., 2002)	96	lung tumors (**86**) normals (**10**)	7129	Affymetrix HuGeneFL	7094
Lung Cancer Subtype Classification: adenocarcinoma vs. squamous cell carcinoma lung tumors	(Bhattacharjee et al., 2001)	160	adenocarcinoma (**139**) squamous (**21**)	12600	Affymetrix U95A	(Su et al., 2001)	28	adenocarcinoma (**14**) squamous (**14**)	12533	Affymetrix U95A	12533
Breast Cancer Subtype Classification: estrogen receptor positive (ER+) vs. ER- breast tumors; untreated patients	(Wang et al., 2005)	286	ER+ (**209**) ER- (**77**)	22283	Affymetrix U133A	(Sotiriou et al., 2006)	119	ER+ (**85**) ER- (**34**)	22283	Affymetrix U133A	22283
Breast Cancer 5 Yr. Prognosis: ER+ patients who developed distant metastases within 5 years (poor prognosis) vs. ones who did not (good prognosis)	(Wang et al., 2005)	204	poor prognosis (**66**) good prognosis (**138**)	22283	Affymetrix U133A	(Sotiriou et al., 2006)	72	poor prognosis (**13**) good prognosis (**59**)	22283	Affymetrix U133A	22283
Glioma Subtype Classification: grade III vs. grade IV glioma tumors	(Phillips et al., 2006)	100	grade III (**24**) grade IV (**76**)	22283	Affymetrix U133A	(Freije et al., 2004)	85	grade III (**26**) grade IV (**59**)	22283	Affymetrix U133A	22283
Leukemia 5 Yr. Prognosis: patients with disease-free survival < 5 years (ones who had relapse or competing events within 5 years) vs. > 5 years	(Yeoh et al., 2002)	164	survival < 5 yr. (**29**) survival > 5 yr. (**135**)	12625	Affymetrix U95A	(Ross et al., 2003)	79	survival < 5 yr. (**18**) survival > 5 yr. (**61**)	22283	Affymetrix U133A	10507

Table 13: Gene expression microarray datasets that were used in independent-dataset experiments.

Lung Cancer Diagnosis

Method to induce multiple signatures	Number of signatures	Number of genes in a signature		Classification performance (AUC)			
				In discovery dataset		In validation dataset	
		mean	95% interval	mean	95% interval	mean	95% interval
TIE*	348/348/187	6	[4 - 8]	0.999	[0.994 - 1.000]	0.998	[0.988 - 1.000]
Resampling+SVM-RFE1	5000/2966/48	9	[1 - 43]	0.987	[0.919 - 1.000]	0.989	[0.949 - 1.000]
Resampling+SVM-RFE2	5000/341/61	1	[1 - 2]	0.967	[0.861 - 1.000]	0.962	[0.633 - 1.000]
Resampling+Univariate1	5000/2199/19	19	[1 - 62]	0.99	[0.919 - 1.000]	0.992	[0.949 - 1.000]
Resampling+Univariate2	5000/294/58	1	[1 - 2]	0.969	[0.861 - 1.000]	0.973	[0.887 - 1.000]
KIAMB1	985/985/985	41	[39 - 42]	0.999	[0.990 - 1.000]	0.995	[0.984 - 1.000]
KIAMB2	1489/1320/1246	48	[12 - 68]	0.999	[0.990 - 1.000]	0.995	[0.978 - 1.000]
KIAMB3	5000/271/157	9	[6 - 15]	0.996	[0.981 - 1.000]	0.997	[0.992 - 1.000]
Iterative Removal	51/51/51	7	[5 - 10]	0.987	[0.919 - 1.000]	0.977	[0.880 - 1.000]

Lung Cancer Subtype Classification

Method to induce multiple signatures	Number of signatures	Number of genes in a signature		Classification performance (AUC)			
				In discovery dataset		In validation dataset	
		mean	95% interval	mean	95% interval	mean	95% interval
TIE*	668/668/413	7	[5 - 8]	0.987	[0.973 - 1.000]	0.973	[0.929 - 1.000]
Resampling+SVM-RFE1	5000/4267/20	392	[1 - 5037]	0.98	[0.909 - 1.000]	0.978	[0.888 - 1.000]
Resampling+SVM-RFE2	5000/1206/107	2	[1 - 5]	0.925	[0.650 - 0.985]	0.914	[0.668 - 0.995]
Resampling+Univariate1	5000/4590/55	528	[1 - 8703]	0.98	[0.903 - 1.000]	0.98	[0.883 - 1.000]
Resampling+Univariate2	5000/917/81	3	[1 - 6]	0.922	[0.839 - 0.988]	0.916	[0.770 - 0.985]
KIAMB1	994/968/965	26	[24 - 26]	0.986	[0.967 - 1.000]	0.982	[0.923 - 1.000]
KIAMB2	1006/1005/1005	48	[47 - 50]	0.99	[0.973 - 1.000]	0.982	[0.923 - 1.000]
KIAMB3	3520/1364/1209	16	[8 - 31]	0.98	[0.948 - 0.997]	0.982	[0.923 - 1.000]
Iterative Removal	29/29/29	8	[5 - 12]	0.978	[0.867 - 1.000]	0.972	[0.882 - 1.000]

Breast Cancer Subtype Classification

Method to induce multiple signatures	Number of signatures	Number of genes in a signature		Classification performance (AUC)			
				In discovery dataset		In validation dataset	
		mean	95% interval	mean	95% interval	mean	95% interval
TIE*	2776/2776/1602	17	[14 - 21]	0.847	[0.824 - 0.873]	0.887	[0.852 - 0.916]
Resampling+SVM-RFE1	5000/4601/22	1627	[1 - 10746]	0.845	[0.821 - 0.888]	0.812	[0.604 - 0.893]
Resampling+SVM-RFE2	5000/2033/65	18	[1 - 135]	0.858	[0.736 - 0.930]	0.761	[0.554 - 0.874]
Resampling+Univariate1	5000/4122/15	3560	[1 - 22283]	0.857	[0.826 - 0.920]	0.823	[0.771 - 0.877]
Resampling+Univariate2	5000/794/22	7	[1 - 18]	0.873	[0.754 - 0.930]	0.814	[0.725 - 0.874]
KIAMB1	983/970/960	31	[30 - 32]	0.85	[0.804 - 0.883]	0.68	[0.427 - 0.846]
KIAMB2	994/964/962	28	[27 - 29]	0.85	[0.802 - 0.884]	0.685	[0.418 - 0.850]
KIAMB3	943/570/493	14	[12 - 15]	0.856	[0.786 - 0.884]	0.694	[0.432 - 0.851]
Iterative Removal	34/34/34	19	[14 - 23]	0.833	[0.793 - 0.866]	0.834	[0.720 - 0.899]

Table 14 (continued on the next page): Results for the number of output signatures (total/unique/unique and non-reducible), number of genes in a signature, and phenotypic classification performance in discovery and validation microarray datasets for independent-dataset experiments. The length of highlighting corresponds to magnitude of the metric (number of genes in a signature or classification performance) relative to other multiple signature extraction methods. The 95% intervals correspond to the observed [2.5 - 97.5] percentile interval over multiple signatures discovered by the method. Uniqueness and non-reducibility of each signature is assessed relative to the corresponding signature extraction method.

Breast Cancer 5 Yr. Prognosis

Method to induce multiple signatures	Number of signatures	Number of genes in a signature		Classification performance (AUC)			
				In discovery dataset		In validation dataset	
		mean	95% interval	mean	95% interval	mean	95% interval
TIE*	5342/5342/3321	84	[81 - 89]	0.671	[0.658 - 0.686]	0.697	[0.674 - 0.720]
Resampling+SVM-RFE1	5000/4755/42	4687	[2 - 22283]	0.684	[0.541 - 0.746]	0.64	[0.487 - 0.752]
Resampling+SVM-RFE2	5000/3407/350	56	[1 - 404]	0.586	[0.413 - 0.719]	0.598	[0.398 - 0.822]
Resampling+Univariate1	5000/4002/29	5791	[1 - 22283]	0.685	[0.573 - 0.741]	0.645	[0.468 - 0.801]
Resampling+Univariate2	5000/2573/139	44	[1 - 162]	0.62	[0.467 - 0.712]	0.628	[0.411 - 0.807]
KIAMB1	986/552/550	14	[14 - 14]	0.596	[0.507 - 0.693]	0.562	[0.399 - 0.716]
KIAMB2	988/969/955	28	[27 - 29]	0.595	[0.482 - 0.708]	0.562	[0.390 - 0.713]
KIAMB3	1182/916/889	23	[12 - 28]	0.596	[0.483 - 0.704]	0.567	[0.394 - 0.724]
Iterative Removal	31/31/31	28	[12 - 82]	0.69	[0.589 - 0.794]	0.606	[0.434 - 0.735]

Glioma Subtype Classification

Method to induce multiple signatures	Number of signatures	Number of genes in a signature		Classification performance (AUC)			
				In discovery dataset		In validation dataset	
		mean	95% interval	mean	95% interval	mean	95% interval
TIE*	5753/5753/4588	46	[45 - 53]	0.871	[0.860 - 0.885]	0.844	[0.830 - 0.860]
Resampling+SVM-RFE1	5000/4255/43	301	[2 - 3599]	0.808	[0.630 - 0.915]	0.74	[0.528 - 0.880]
Resampling+SVM-RFE2	5000/2055/126	3	[1 - 13]	0.694	[0.545 - 0.890]	0.637	[0.463 - 0.830]
Resampling+Univariate1	5000/4751/63	925	[2 - 17022]	0.84	[0.690 - 0.905]	0.818	[0.554 - 0.919]
Resampling+Univariate2	5000/1926/117	3	[1 - 15]	0.74	[0.495 - 0.900]	0.65	[0.450 - 0.860]
KIAMB1	973/658/654	15	[15 - 15]	0.765	[0.675 - 0.865]	0.71	[0.558 - 0.811]
KIAMB2	974/964/964	30	[29 - 30]	0.781	[0.685 - 0.880]	0.732	[0.610 - 0.832]
KIAMB3	1408/786/746	21	[6 - 30]	0.77	[0.685 - 0.865]	0.728	[0.588 - 0.821]
Iterative Removal	58/58/58	24	[15 - 44]	0.847	[0.744 - 0.921]	0.842	[0.743 - 0.914]

Leukemia 5 Yr. Prognosis

Method to induce multiple signatures	Number of signatures	Number of genes in a signature		Classification performance (AUC)			
				In discovery dataset		In validation dataset	
		mean	95% interval	mean	95% interval	mean	95% interval
TIE*	1804/1804/1561	22	[20 - 28]	0.714	[0.647 - 0.805]	0.711	[0.631 - 0.784]
Resampling+SVM-RFE1	5000/4643/158	1984	[1 - 8756]	0.631	[0.422 - 0.741]	0.612	[0.440 - 0.725]
Resampling+SVM-RFE2	5000/2537/570	15	[1 - 92]	0.543	[0.341 - 0.749]	0.55	[0.356 - 0.725]
Resampling+Univariate1	5000/3897/116	4024	[1 - 10507]	0.649	[0.431 - 0.756]	0.606	[0.419 - 0.717]
Resampling+Univariate2	5000/2516/465	48	[1 - 329]	0.539	[0.235 - 0.756]	0.529	[0.342 - 0.725]
KIAMB1	988/984/984	31	[29 - 31]	0.515	[0.351 - 0.681]	0.603	[0.445 - 0.735]
KIAMB2	1213/1131/1127	46	[13 - 56]	0.517	[0.341 - 0.687]	0.602	[0.460 - 0.732]
KIAMB3	4485/30/30	7	[6 - 10]	0.438	[0.348 - 0.632]	0.563	[0.530 - 0.760]
Iterative Removal	2/2/2	21	[19 - 23]	0.673	[0.630 - 0.716]	0.652	[0.550 - 0.753]

Table 14 (continued from the previous page)

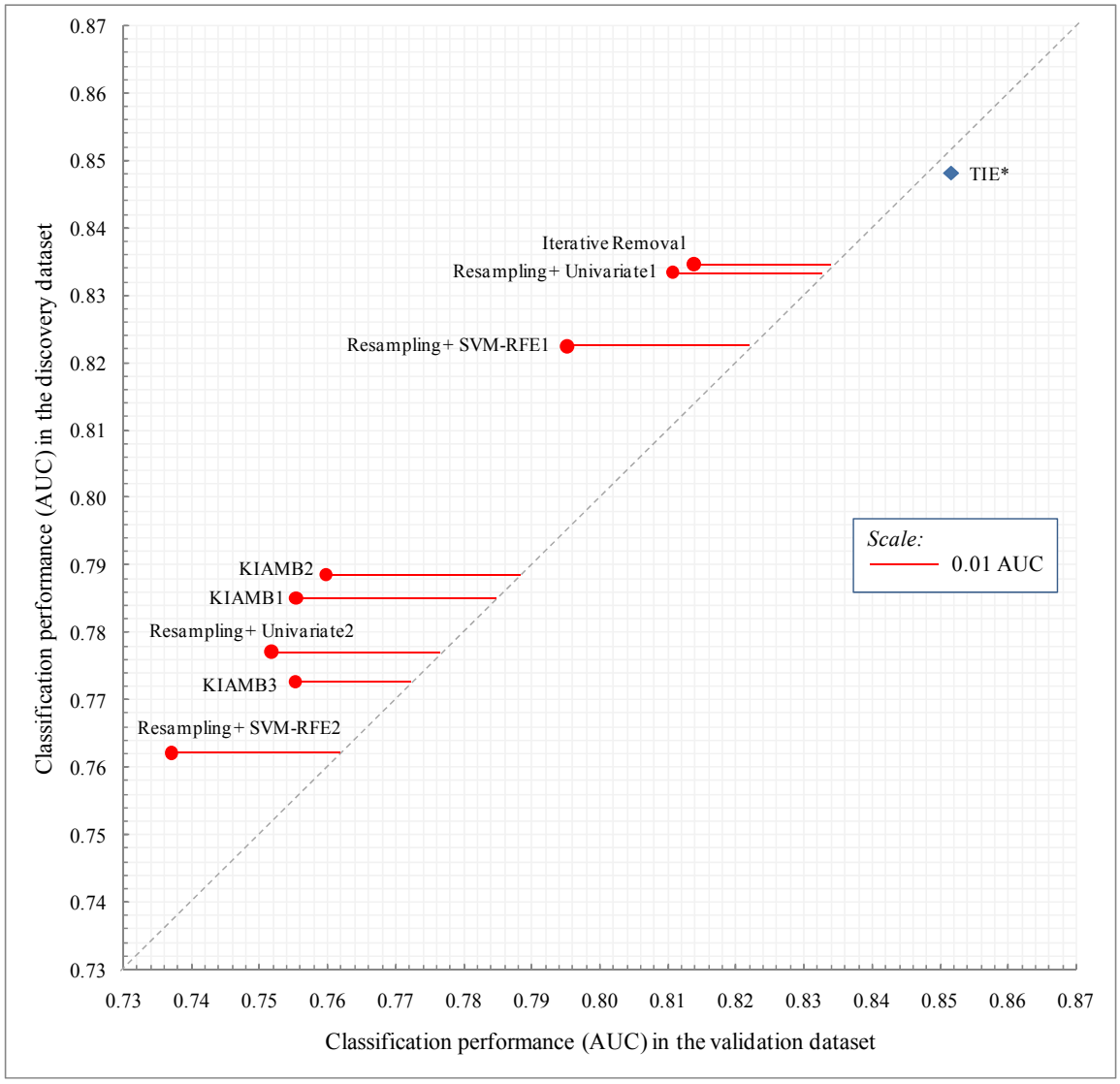


Figure 15: Plot of classification performance (AUC) in the validation dataset versus classification performance in the discovery dataset averaged over 6 pairs of microarray gene expression datasets. Axes are magnified for better visualization. The classification performance of a signature produced by HITON-PC (which is included in the output of TIE*) is very similar to an average signature produced by TIE*. Specifically, the performance of HITON-PC signature in discovery and validation data is 0.850 and 0.860 AUC, respectively. The performance of an average TIE* signature in discovery and validation data is 0.848 and 0.850, respectively.

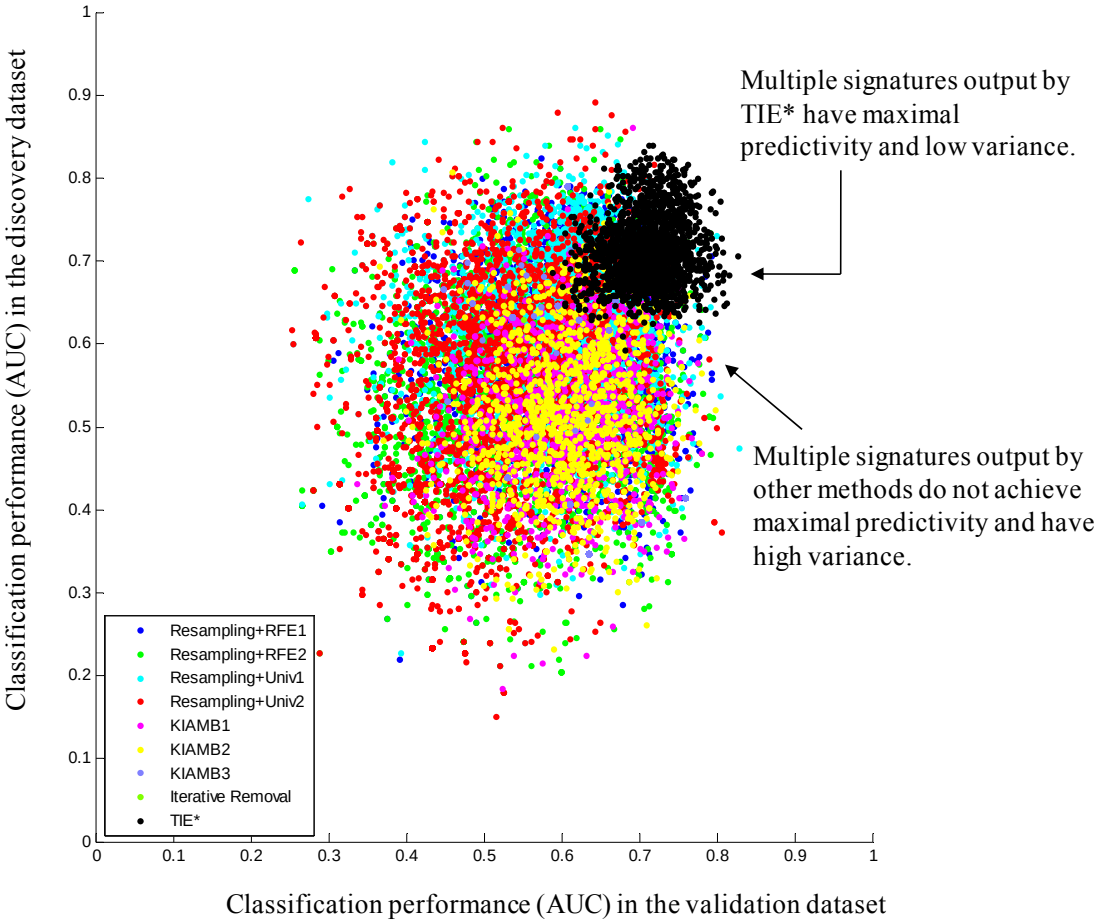


Figure 16: Plot of classification performance (AUC) in the validation dataset versus classification performance in the discovery dataset for each signature output by each method for the *Leukemia 5 yr. Prognosis* task. Each dot in the graph corresponds to a signature (SVM computational model of the phenotype).

Task	Number of genes common in X% of discovered signatures					
	50%	60%	70%	80%	90%	100%
Lung Cancer Diagnosis	4	0	0	0	0	0
Lung Cancer Subtype Classification	5	0	0	0	0	0
Breast Cancer Subtype Classification	15	11	4	1	0	0
Breast Cancer 5 Yr. Prognosis	85	85	84	84	84	1
Glioma Subtype Classification	48	48	48	47	41	0
Leukemia 5 Yr. Prognosis	23	23	23	20	1	0

Table 15: Number of common genes in 50%, 60%, …, 100% of signatures discovered by TIE* algorithm for each dataset.

Eight state-of-the-art algorithms for multiple signature extraction were used to compare to TIE* as described in Appendix G.

The experiments first involved running TIE* and baseline comparison algorithms on discovery datasets to identify all maximally predictive and non-redundant signatures and estimate their predictivity by holdout validation. Then reproducibility of all identified signatures was assessed in independent validation datasets. A linear SVM classifier (Vapnik, 1998) was used in all experiments to build signatures from the selected genes. The predictive performance of resulting signatures was measured by area under ROC curve (AUC) metric (Fawcett, 2003). Statistical comparisons of predictivity between methods in the same dataset were accomplished by Wilcoxon rank sum test with $\alpha = 0.05$ (Hollander and Wolfe, 1999).

The detailed results of experiments are provided in Table 14. As can be seen, TIE* achieves maximal classification performance in 5 out of 6 validation datasets. Non-TIE* methods achieve maximal classification performance in 0 to 2 datasets depending on the method. In the dataset where TIE* has predictivity that is statistically distinguishable from the empirical maximal (*Lung Cancer Subtype Classification*), the magnitude of this difference is <0.009 AUC on average over all discovered signatures, thus this particular deviation from maximal predictivity may be considered negligible for most practical purposes.

Figure 15 plots predictivity estimated in the discovery dataset (using an unbiased error estimator and protocol) against predictivity verified in the validation dataset for each method averaged over all datasets and all discovered signatures. Recall that validation datasets originate from different laboratories and/or using different microarray platforms than discovery datasets. The horizontal distance of each method to the diagonal measures the magnitude of overfitting defined as the difference (ε_1-ε_2), where ε_1 = expected performance in the validation data obtained by holdout validation in the discovery dataset, and ε_2 = observed validation dataset performance. TIE* rests slightly right of the diagonal denoting no overfitting, or equivalently perfect statistical reproducibility on average. However all other methods exhibit varying degrees of non-reproducibility. Depending on method the average magnitude of overfitting varies from 0.02 to 0.03 AUC.

Figure 16 plots predictivity in the validation dataset versus predictivity in the discovery dataset for each signature output by each method for the *Leukemia 5 yr. Prognosis* task. As can be seen, multiple signatures output by TIE* have maximal predictivity both in the discovery and validation datasets and low variance. On the other hand, multiple signatures output by other methods typically have lower predictivity and/or high variance. Similar trends can be also observed in other datasets.

Finally, analysis of the signatures output by TIE* reveals that they share many genes in common. Table 15 shows the number of common genes in 50%, 60%, …, 100% of output signatures for each dataset. Genes differ in the percentage of signatures they participate in. A heuristic that genes that belong to the higher fractions of signatures are localized closer to the pathway(s) affecting and being affected by the phenotypic response variable may be useful in exploratory studies, however this does not hold in all distributions (Aliferis et al., 2006b).

Single-dataset experiments

The experiments reported in this section are primarily concerned with an additional evaluation of TIE* and baseline algorithms for multiple signature extraction in terms of maximal predictivity in datasets with relatively large sample size.

Seven human gene expression microarray datasets used in the experiments are described in Table 16. None of these datasets was used in experiments from the previous section. The following experimental design was adopted: A large portion of the dataset (with >100 samples, referred to as "*discovery dataset*") was used for signature extraction and performance estimation by holdout validation and another non-overlapping large portion (with 100 samples, referred to as "*validation dataset*") was used for an additional performance estimation. To minimize variance due to splitting of the data into non-overlapping discovery and validation datasets, I performed 10 balanced splits of the data and ran all algorithms on each split. Therefore, the experiments were 10 times more computationally expensive than the independent-dataset evaluation.

The following instantiation of the TIE* algorithm was used in experiments. It can be described by a tuple of input components ($\mathbb{X}$, $\mathbb{Y}$, $\mathbb{Z}$):

- $\mathbb{X}$ (Markov boundary algorithm) = HITON-PC (Figure 9) that uses Fisher's Z test with α = 0.05;
- $\mathbb{Y}$ (strategy to generate subsets of variables that have to be removed to identify new Markov boundaries of *T*) = *IncLex* (Figure 13) with the maximum size of subset **G** limited to 5 genes.
- $\mathbb{Z}$ (criterion to verify Markov boundaries) = *Predictivity* (Figure 12) that uses:
    - $\mathbb{L}$ = linear SVM classifier (Vapnik, 1998);
    - $\mathbb{M}$ = area under ROC curve (AUC) performance metric (Fawcett, 2003);
    - $\mathbb{E}$ = holdout validation performance estimator;

- $\mathbb{C}$ = nonparametric method to compare estimates of AUC with $\alpha = 0.1$ (DeLong et al., 1988).

Eight state-of-the-art algorithms for multiple signature extraction were used to compare to TIE* as described in Appendix G.

The experiments first involved running TIE* and baseline comparison algorithms on discovery datasets to identify all maximally predictive and non-redundant signatures and estimate their predictivity by holdout validation. Then, reproducibility of all identified signatures was assessed in non-overlapping validation datasets. A linear SVM classifier (Vapnik, 1998) was used in all experiments to build signatures from the selected genes. The predictive performance of resulting signatures was measured by area under ROC curve (AUC) metric (Fawcett, 2003). Statistical comparisons of predictivity between methods in the same dataset were accomplished by Wilcoxon rank sum test with $\alpha = 0.05$ (Hollander and Wolfe, 1999).

The detailed results of experiments are provided in Table 17. It is worth noting that TIE* achieves maximal classification performance in 6 out of 7 validation datasets while non-TIE* methods achieve maximal classification performance in 0 to 1 datasets depending on the method. In the dataset where TIE* has predictivity that is statistically distinguishable from the empirical maximal (*Breast Cancer Subtype Classification II*), the magnitude of this difference is <0.01 AUC on average over all discovered signatures, which can be considered negligible for most practical purposes.

Task	Reference	Sample size	Samples per class	Number of genes	Microarray platform
Lymphoma Subtype Classification I: Diffuse large-B-cell lymphoma (DLBCL) vs. Burkitt's lymphoma (BL) patients	(Dave et al., 2006)	303	DLBCL (**258**) BL (**45**)	2745	Human LymphDx 2.7k GeneChip
Lymphoma Subtype Classification II: Diffuse large-B-cell lymphoma (DLBCL) vs. mediastinal large B-cell (MLBCL) patients	(Savage et al., 2003)	210	DLBCL (**176**) MLBCL (**34**)	32403 (44928)	Affymetrix U133A and U133B
Breast Cancer Subtype Classification I: p53 mutant vs. wild-type breast tumors	(Miller et al., 2005)	251	p53 mutant (**58**) p53 wild-type (**193**)	22283	Affymetrix U133A
Breast Cancer Subtype Classification II: estrogen receptor positive (ER+) vs. ER- breast tumors	(Miller et al., 2005)	247	ER+ (**213**) ER- (**34**)	22283	Affymetrix U133A
Breast Cancer Subtype Classification III: progesterone receptor positive (PgR+) vs. PgR- breast tumors	(Miller et al., 2005)	251	PgR+ (**190**) PgR- (**61**)	22283	Affymetrix U133A
Breast Cancer 5 Yr. Prognosis: ER+ patients who developed distant metastases within 5 years (poor prognosis) vs. ones who did not (good prognosis)	(van de Vijver et al., 2002)	215	poor prognosis (**51**) good prognosis (**164**)	24496	Agilent Hu25K
Bladder Cancer Stage Classification: stage Ta. vs. other stages (T1, T2, T3, T4) of bladder tumors	(Dyrskjot et al., 2007)	404	stage Ta (**189**) other stages (**215**)	1381 (3072)	MDL Human 3k

Table 16: Gene expression microarray datasets that were used in single-dataset experiments. For the task of *Lymphoma Subtype Classification II*, a version of this dataset with 32,403 genes (obtained by excluding gene probes absent in all samples) is used. For the *Bladder Cancer Stage Classification* task, a version of this dataset processed by its authors with only 1,381 genes is used.

Lymphoma subtype classification I

Method to induce multiple signatures	Number of signatures	Number of genes in a signature		Classification performance (AUC)			
				In discovery dataset		In validation dataset	
		mean	95% interval	mean	95% interval	mean	95% interval
TIE*	2767/2767/1439	10	[8 - 13]	0.992	[0.982 - 0.999]	0.983	[0.971 - 0.992]
Resampling+SVM-RFE1	5000/4012/58	65	[1 - 495]	0.987	[0.954 - 0.999]	0.974	[0.925 - 0.993]
Resampling+SVM-RFE2	5000/1117/82	2	[1 - 5]	0.957	[0.839 - 0.995]	0.934	[0.827 - 0.985]
Resampling+Univariate1	5000/3476/22	168	[2 - 1223]	0.988	[0.970 - 0.997]	0.972	[0.949 - 0.987]
Resampling+Univariate2	5000/536/36	1	[1 - 3]	0.971	[0.910 - 0.993]	0.949	[0.888 - 0.984]
KIAMB1	1129/1107/1088	73	[23 - 82]	0.993	[0.982 - 0.999]	0.983	[0.970 - 0.992]
KIAMB2	5000/2860/2587	26	[12 - 72]	0.992	[0.980 - 0.999]	0.98	[0.966 - 0.991]
KIAMB3	5000/274/212	9	[7 - 13]	0.991	[0.980 - 0.999]	0.978	[0.965 - 0.989]
Iterative Removal	30/30/30	10	[7 - 13]	0.987	[0.967 - 0.998]	0.974	[0.949 - 0.991]

Lymphoma subtype classification II

Method to induce multiple signatures	Number of signatures	Number of genes in a signature		Classification performance (AUC)			
				In discovery dataset		In validation dataset	
		mean	95% interval	mean	95% interval	mean	95% interval
TIE*	5140/5140/3399	18	[15 - 22]	0.818	[0.738 - 0.881]	0.791	[0.738 - 0.833]
Resampling+SVM-RFE1	5000/4756/82	2696	[1 - 19554]	0.821	[0.577 - 0.928]	0.79	[0.579 - 0.887]
Resampling+SVM-RFE2	5000/2862/371	13	[1 - 59]	0.669	[0.357 - 0.900]	0.655	[0.390 - 0.846]
Resampling+Univariate1	5000/3464/55	6635	[2 - 31863]	0.811	[0.605 - 0.919]	0.785	[0.611 - 0.885]
Resampling+Univariate2	5000/2068/231	84	[1 - 221]	0.682	[0.379 - 0.884]	0.679	[0.403 - 0.856]
KIAMB1	977/895/870	20	[19 - 20]	0.751	[0.561 - 0.906]	0.747	[0.631 - 0.843]
KIAMB2	973/724/706	17	[16 - 18]	0.75	[0.564 - 0.899]	0.749	[0.633 - 0.844]
KIAMB3	1188/309/296	10	[6 - 13]	0.753	[0.560 - 0.896]	0.753	[0.624 - 0.840]
Iterative Removal	17/17/17	13	[9 - 20]	0.788	[0.667 - 0.890]	0.729	[0.641 - 0.844]

Breast cancer subtype classification I

Method to induce multiple signatures	Number of signatures	Number of genes in a signature		Classification performance (AUC)			
				In discovery dataset		In validation dataset	
		mean	95% interval	mean	95% interval	mean	95% interval
TIE*	4343/4343/4250	91	[90 - 98]	0.871	[0.862 - 0.882]	0.857	[0.850 - 0.865]
Resampling+SVM-RFE1	5000/4776/63	1140	[1 - 9974]	0.843	[0.693 - 0.917]	0.823	[0.703 - 0.879]
Resampling+SVM-RFE2	5000/2972/212	10	[1 - 67]	0.747	[0.497 - 0.905]	0.731	[0.521 - 0.873]
Resampling+Univariate1	5000/4475/64	2449	[1 - 19192]	0.857	[0.737 - 0.917]	0.837	[0.745 - 0.879]
Resampling+Univariate2	5000/2063/132	10	[1 - 46]	0.775	[0.592 - 0.907]	0.764	[0.598 - 0.875]
KIAMB1	977/902/899	29	[28 - 29]	0.76	[0.639 - 0.861]	0.756	[0.661 - 0.833]
KIAMB2	980/913/910	28	[27 - 28]	0.76	[0.641 - 0.863]	0.755	[0.662 - 0.834]
KIAMB3	1012/590/569	17	[10 - 18]	0.755	[0.646 - 0.852]	0.754	[0.665 - 0.826]
Iterative Removal	24/24/24	56	[38 - 97]	0.867	[0.807 - 0.916]	0.838	[0.788 - 0.879]

Table 17 (continued on the next page): Results for the number of output signatures (total/unique/unique and non-reducible), number of genes in a signature, and phenotypic classification performance in discovery and validation microarray datasets for single-dataset experiments. The length of highlighting corresponds to magnitude of the metric (number of genes in a signature or classification performance) relative to other multiple signature extraction methods. The 95% intervals correspond to the observed [2.5 - 97.5] percentile interval over multiple signatures discovered by the method. Uniqueness and non-reducibility of each signature is assessed relative to the corresponding signature extraction method.

Breast cancer subtype classification II

Method to induce multiple signatures	Number of signatures	Number of genes in a signature		Classification performance (AUC)			
				In discovery dataset		In validation dataset	
		mean	95% interval	mean	95% interval	mean	95% interval
TIE*	4312/4312/2718	15	[11 - 20]	0.902	[0.858 - 0.939]	0.858	[0.819 - 0.894]
Resampling+SVM-RFE1	5000/4289/54	479	[1 - 4090]	0.89	[0.766 - 0.947]	0.854	[0.747 - 0.904]
Resampling+SVM-RFE2	5000/2341/224	6	[1 - 38]	0.816	[0.501 - 0.946]	0.784	[0.501 - 0.906]
Resampling+Univariate1	5000/3856/43	833	[1 - 11196]	0.911	[0.812 - 0.950]	0.868	[0.788 - 0.908]
Resampling+Univariate2	5000/1460/125	3	[1 - 13]	0.857	[0.648 - 0.949]	0.814	[0.607 - 0.903]
KIAMB1	982/978/973	31	[31 - 32]	0.837	[0.701 - 0.934]	0.814	[0.696 - 0.888]
KIAMB2	980/972/967	31	[30 - 31]	0.838	[0.706 - 0.932]	0.814	[0.698 - 0.889]
KIAMB3	997/564/543	15	[10 - 16]	0.833	[0.710 - 0.930]	0.816	[0.700 - 0.886]
Iterative Removal	26/26/26	14	[10 - 21]	0.896	[0.811 - 0.942]	0.852	[0.787 - 0.899]

Breast cancer subtype classification III

Method to induce multiple signatures	Number of signatures	Number of genes in a signature		Classification performance (AUC)			
				In discovery dataset		In validation dataset	
		mean	95% interval	mean	95% interval	mean	95% interval
TIE*	6306/6306/4638	75	[73 - 79]	0.809	[0.794 - 0.824]	0.79	[0.781 - 0.798]
Resampling+SVM-RFE1	5000/4733/93	1025	[1 - 8298]	0.765	[0.583 - 0.854]	0.74	[0.590 - 0.808]
Resampling+SVM-RFE2	5000/2570/436	10	[1 - 83]	0.656	[0.418 - 0.831]	0.641	[0.447 - 0.784]
Resampling+Univariate1	5000/4424/73	2265	[1 - 19448]	0.785	[0.645 - 0.862]	0.753	[0.629 - 0.809]
Resampling+Univariate2	5000/2056/270	8	[1 - 27]	0.694	[0.476 - 0.843]	0.669	[0.495 - 0.786]
KIAMB1	984/977/974	30	[29 - 30]	0.713	[0.595 - 0.826]	0.698	[0.597 - 0.788]
KIAMB2	976/799/793	24	[23 - 24]	0.708	[0.589 - 0.821]	0.697	[0.601 - 0.781]
KIAMB3	1071/616/595	19	[10 - 22]	0.71	[0.595 - 0.819]	0.693	[0.597 - 0.778]
Iterative Removal	38/38/38	35	[19 - 71]	0.81	[0.741 - 0.870]	0.774	[0.714 - 0.824]

Breast cancer 5 yr. prognosis

Method to induce multiple signatures	Number of signatures	Number of genes in a signature		Classification performance (AUC)			
				In discovery dataset		In validation dataset	
		mean	95% interval	mean	95% interval	mean	95% interval
TIE*	5800/5800/4999	39	[37 - 43]	0.65	[0.612 - 0.689]	0.72	[0.694 - 0.750]
Resampling+SVM-RFE1	5000/4675/132	962	[1 - 9727]	0.621	[0.421 - 0.756]	0.682	[0.518 - 0.779]
Resampling+SVM-RFE2	5000/2369/573	5	[1 - 29]	0.561	[0.333 - 0.771]	0.59	[0.405 - 0.743]
Resampling+Univariate1	5000/3801/81	1684	[1 - 18370]	0.626	[0.426 - 0.758]	0.697	[0.535 - 0.790]
Resampling+Univariate2	5000/1876/366	7	[1 - 29]	0.563	[0.349 - 0.771]	0.61	[0.441 - 0.753]
KIAMB1	980/967/963	27	[26 - 27]	0.596	[0.427 - 0.757]	0.62	[0.516 - 0.730]
KIAMB2	979/775/764	20	[14 - 21]	0.592	[0.424 - 0.751]	0.617	[0.515 - 0.729]
KIAMB3	2891/261/237	9	[4 - 19]	0.593	[0.443 - 0.724]	0.611	[0.529 - 0.715]
Iterative Removal	68/68/68	19	[12 - 37]	0.664	[0.537 - 0.787]	0.69	[0.603 - 0.766]

Bladder cancer stage classification

Method to induce multiple signatures	Number of signatures	Number of genes in a signature		Classification performance (AUC)			
				In discovery dataset		In validation dataset	
		mean	95% interval	mean	95% interval	mean	95% interval
TIE*	5125/5125/4550	34	[32 - 39]	0.831	[0.823 - 0.840]	0.823	[0.815 - 0.830]
Resampling+SVM-RFE1	5000/4555/88	281	[2 - 1293]	0.793	[0.698 - 0.837]	0.792	[0.702 - 0.830]
Resampling+SVM-RFE2	5000/2688/99	12	[1 - 88]	0.727	[0.587 - 0.825]	0.728	[0.581 - 0.818]
Resampling+Univariate1	5000/4104/22	181	[2 - 1037]	0.799	[0.747 - 0.838]	0.799	[0.750 - 0.831]
Resampling+Univariate2	5000/1219/20	4	[1 - 10]	0.757	[0.674 - 0.826]	0.759	[0.679 - 0.819]
KIAMB1	5000/291/220	10	[6 - 16]	0.794	[0.752 - 0.829]	0.791	[0.756 - 0.820]
KIAMB2	5000/85/64	6	[4 - 9]	0.793	[0.760 - 0.817]	0.79	[0.762 - 0.812]
KIAMB3	5000/22/17	3	[3 - 4]	0.783	[0.759 - 0.804]	0.777	[0.759 - 0.803]
Iterative Removal	38081	35	[30 - 40]	0.819	[0.808 - 0.833]	0.807	[0.790 - 0.823]

Table 17 (continued from the previous page)

CHAPTER VIII

DISCUSSION

On related methods from the field of statistics

The present section provides an overview of related methods from the field of statistics. The methods listed below were not used in numerical experiments of the present thesis because they do not output multiple Markov boundaries and are not designed to do so.

The discipline of classical statistics offers several methods for diagnostics of regression and generalized linear models and identification of the sources of multicollinearity. Multicollinearity occurs when there is a linear relationship among some of the predictor variables in the data. This in turn can lead to existence of multiple Markov boundaries. Notable works in the field are (Belsley et al., 1980), (Stewart, 1987), (Hadi and Velleman, 1987), and (Weissfeld and Sereika, 1991). These methods typically build on the observation that small eigenvalues of the cross-products data matrix $\mathbf{X}^T\mathbf{X}$ indicate multicollinearity. As far as the problem of identification of multiple Markov boundaries is concerned, an obvious shortcoming of this methodology is inability to detect cases when variables are not multicollinear (or nearly multicollinear) but still provide equivalent information about the response variable (e.g., variables *A* and *B* in Figure 1). In addition, the above methods cannot detect nonlinear relations among predictor variables.

Several researchers propose to use clustering techniques to identify groups of highly correlated variables. The works (Meinshausen, 2008), (Park et al., 2007), and (Hastie et al., 2001) apply unsupervised hierarchical clustering methods to identify variables that are highly correlated. On the other hand, the works (Hastie et al., 2000), (Jornsten and Yu, 2003), (Dettling and Buhlmann, 2004), and (Dettling and Buhlmann, 2002) propose a solution to the similar

problem using supervised clustering techniques that take into account information about the response variable. The methods based on unsupervised clustering besides having other limitations will fail to group variables that are similar only when the response variable is considered (as are *A* and *B* in Figure 1). The methods based on supervised clustering are typically complex multi-stage algorithms that are heuristic and sometimes use unsupervised methods (e.g., k-means clustering, PCA) in a semi-supervised fashion.

Most recent research proposes to use objective functions in the statement of regression/classification problems that will assign the same coefficients to highly correlated variables that are important for prediction of the response variable. The work (Zou and Hastie, 2005) proposes LARS-EN regression algorithm that uses an L_2-norm loss and elastic net penalty which is a mixture of the L_1 and L_2-norm penalties. Similarly, (Wang et al., 2006) proposes the DrSVM classification algorithm that uses a hinge loss function and elastic net penalty. The elastic net penalty allows to obtain sparse solutions by setting to zero coefficients of predictor variables that are not relevant for prediction of the response variable (which is a property of the L_1-norm penalty). At the same time, it encourages to select (or remove) together highly correlated variables (which is a property of the L_2-norm penalty). The work (Bondell and Reich, 2008) introduces the OSCAR regression algorithm that uses an L_2-norm loss and a penalty that is a mixture of the L_1 and pairwise L_∞-norms. Again, the L_1-norm encourages sparseness of solutions and pairwise L_∞-norm encourages equality of coefficients. As it is illustrated in (Bondell and Reich, 2008), the grouping property of OSCAR penalty in much stronger than that of elastic net penalty.

In some restricted distributions the algorithms LARS-EN, DrSVM, and OSCAR can identify members of multiple Markov boundaries (by assigning them nonzero coefficients) and provide information on how specifically to construct multiple Markov boundaries (by assigning the same coefficient to variables that are interchangeable for maximal prediction of the response variable). However, this is not the case in general, and there are many situations when the above

algorithms will fail. For example, consider a generative model with response variable Y and predictor variables X_1, …, X_{10} that are distributed as $\mathbb{N}(0,1)$. All variables except for X_3 are generated at random, and $X_3 = 1/\sqrt{13}\ (3X_1 + 2X_2)$. The response variable is defined as $Y = X_3 + \varepsilon$, where ε is distributed as $\mathbb{N}(0,0.025)$. There are 2 Markov boundaries of Y in this distribution: $\{X_3\}$ and $\{X_1, X_2\}$. A sample of the size 10,000 was generated from this distribution, and LARS-EN and OSCAR algorithms were applied to it. Indeed, both algorithms assign 0 coefficients to variables X_4, …, X_{10} that do not participate in Markov boundaries. However, OSCAR assigns nonzero coefficient only for variable X_3 and variables X_1 and X_2 receive 0 coefficients. Thus, the algorithm implies that X_1 and X_2 do not participate in a Markov boundary. On the other hand, LARS-EN assigns nonzero coefficients to all variables X_1, X_2, and X_3. However, the magnitudes of these coefficients are different (they are 0.623, 0.417, 0.750 for X_1, X_2, X_3, respectively), thus it is not possible to construct multiple Markov boundaries from the output of this algorithm. In general, the expressivity of an algorithm that outputs a coefficient for each predictor variable is not sufficient to provide information on how to construct multiple Markov boundaries.

What are the factors contributing to molecular signature multiplicity?

The results of this thesis refute or suggest that modifications are needed to several widespread positions about signature multiplicity. For example, the model in Figure 1 demonstrates that signature reproducibility neither precludes multiplicity nor requires sample sizes with thousands of subjects. It also shows that that multiplicity of signatures does not require dense connectivity. Similarly, it shows that noisy measurements or normalization are not necessary conditions for signature multiplicity. The resimulation experiment suggests that networks modeled after real microarray data can exhibit signature multiplicity even in large sample sizes and that in this type of data signature multiplicity is produced by a combination of small sample size-related variance *and* intrinsic multiplicity in the underlying network (due to gene-gene and gene-phenotype relations). The results with real human microarray datasets show

that multiple signatures output by TIE* are reproducible even though they are derived from small sample, noisy, and heavily-processed data.

Overall, the results of this work are consistent with the hypothesis that signature multiplicity in real-life microarray gene expression datasets is created by a combination of several factors that include the following:

1. intrinsic information redundancy due to gene-gene and gene-phenotype relations (Dougherty and Brun, 2006);
2. variability in the output of gene selection and classification algorithms especially in small sample sizes;
3. small sample statistical indistinguishability of signatures with different large sample predictivity and/or redundancy characteristics (e.g., see Appendix J);
4. presence of the hidden/unobserved variables (e.g., see Appendix K);
5. correlated measurement noise components that introduce a bias in gene expression profiles (e.g., noise that is localized in regions of microarray chips) (Balazsi and Oltvai, 2007);
6. RNA amplification techniques that systematically distort measurements of transcript ratios (e.g., double-round T7-based amplification protocol) (Wagner and Radelof, 2007);
7. cellular aggregation and sampling from mixtures of distributions that affect inference of conditional independence relations and thus decisions about redundancy characteristics of the signatures (Chu et al., 2003);
8. normalization and other data pre-processing methods that artificially increase correlations among genes (e.g., multivariate normalization in microarrays) (Qiu et al., 2005; Gold et al., 2005; Ploner et al., 2005);
9. engineered redundancy in the assay technology platforms (e.g., multiple probes for the same gene).

Analysis of multiple signature extraction methods

The signature multiplicity discovery problem is by its nature a combinatorial one and worst-case exponential since distributions exist where the number of maximally predictive and non-redundant signatures is exponential to the number of variables (see Theorem 4). Thus any correct algorithm that finds all such signatures will be also worst-case exponential. A more practical consideration is the *average-case* performance of a sound algorithm *in real data*. In the experiments with resimulated and real human gene expression microarray data, TIE* was run efficiently by constraining the cardinality of subset **G** in line 4 of the algorithm (Figure 4) trading off completeness for execution speed. For example it takes TIE* <1 minute in artificial simulated dataset *TIED2* with 1,000 variables to extract all signatures and up to several hours for real gene expression data using a single Intel Xeon 2.4 GHz CPU.

With regard to non-TIE* baseline comparison algorithms, I note that resampling-based methods that use bootstrap samples to extract signatures may stop producing multiple signatures in large sample sizes. This is expected because resampling methods are designed to address directly only the small sample multiplicity and not the intrinsic multiplicity which persists in large samples. Iterative removal, on the other hand, by its design always fails to identify all maximally predictive and non-redundant signatures that have genes in common. KIAMB among the baseline algorithms has the strongest theoretical motivation because it was shown to discover all Markov boundaries for specific but not all distributions. However, the algorithm exhibits several limitations. A major limitation of KIAMB is that it has sample size requirements that range from at least linear to exponential to the number of genes in a signature (depending on test of independence employed). This makes the algorithm not only computationally inefficient but also prone to statistical errors in small sample sizes. This leads to its substantial observed overfitting in the independent-dataset experiments with real data and inability to find the maximally predictive and non-redundant signatures in simulated data. KIAMB, being a randomized search algorithm, also guarantees to output all signatures that satisfy its distributional

requirements, but only after an infinite number of runs in the worst-case. The method by design will discover the same signatures over and over again further compounding its computational inefficiency.

In molecular high-throughput datasets produced by dissimilar underlying biological mechanisms, assayed with different platforms and pre-processed and modeled with a variety of algorithms, the relative contributions of the factors contributing to signature multiplicity will vary. As a result, methods that rely on a specific cause of multiplicity or combination of causes will not output all maximally predictive and non-redundant signatures in all types of high-throughput data.

Dealing with molecular signature multiplicity using a Markov boundary framework and the TIE* algorithm does not require a particular combination of factors causing signature multiplicity in order to be able to discover all maximally predictive and non-redundant signatures. Because of efficient heuristics TIE* can extract the signature set very efficiently when the connectivity is locally sparse, and the number of true optimal signatures is low-order polynomial or smaller to the number of variables. A very important factor for performance of TIE* is the choice of a Markov boundary algorithm to discover non-redundant and maximally predictive signatures in the distribution at hand. Latest developments in Markov boundary discovery provide such tools for high-throughput data. One of the key advantages of these methods is ability to implicitly control for false discovery rate (Aliferis et al., 2008a; Aliferis et al., 2008b).

Directions for future research

The experiments used real data exclusively from human cancer gene expression microarray datasets because of pragmatic reasons: known identity of observed variables, number and size of available datasets, and maturity of standardization protocols that allow for multiple independent-dataset validation experiments. The methods introduced here are in principle directly applicable to any type of data and problem domain, and future research in this direction is

warranted. The successful results of application of several Markov boundary techniques used in this work to numerous problems outside development of molecular signatures (e.g., information retrieval, predicting bankruptcy, drug discovery, image recognition, ecological modeling) promise very broad applicability of the TIE* algorithm.

Another interesting direction for future research is development of multiple signature extraction algorithms for special distributions. Consider Figure 2 and assume that there are 3 groups of variables with 1000 variables in each (i.e., $m = 3$, $n = 3000$). Thus, there are 1000^3 maximally predictive and non-redundant signatures (Markov boundaries) in this distribution. TIE* would discover all of them, however, it will take very long time. A more efficient solution approach is to learn a single Markov boundary, use statistical methods from the section "On related methods from the field of statistics" to group variables into three clusters/groups, and then simply enumerate all remaining 1000^3-1 Markov boundaries. Even though the above approach provides significant computational savings for such distribution, it will not work in general, e.g. because variables in the cluster may not be members of a Markov boundary.

Conclusion

The contributions of this thesis are four-fold: *First*, I developed a Markov boundary characterization of molecular signature multiplicity. *Second*, I designed a generative algorithm (termed TIE*) that can correctly identify all Markov boundaries (and by extension all maximally predictive and non-redundant molecular signatures) independent of data distribution. The generative algorithm is provably correct given admissible input components and can be instantiated in many ways. *Third*, I conducted an empirical evaluation of the novel algorithm and compared it to existing state-of-the-art methods. Three sources of data were used for this evaluation: artificial simulated data where all maximally predictive and non-redundant signatures are known a priori, resimulated microarray gene expression data, and real human microarray gene expression data. The TIE* algorithm demonstrated excellent empirical performance: it identified

exactly the set of true signatures in artificial datasets, and its signatures have superb predictivity and reproducibility in real human gene expression data. On the other hand, baseline comparison methods either fail to extract most of true signatures in artificial data or incur large number of false positive variables in the discovered signatures. In experiments with real gene expression data, baseline comparison methods either output non-reproducible signatures or signatures with inferior predictivity compared to TIE*. Finally, in experiments with resimulated microarray gene expression data, TIE* discovered the overwhelming majority of maximally predictive and non-redundant signatures output by other methods, thus demonstrating that other techniques typically have very little (if any) contribution to the signatures output by TIE*. *Fourth*, I tested several hypotheses about the causes of molecular signature multiplicity. This led to refinement of several wide-spread hypotheses about this phenomenon.

APPENDIX A

NOTATION AND KEY DEFINITIONS FROM THE THEORY OF LEARNING GRAPHICAL STRUCTURES

In this thesis upper-case letters in italics denote random variables (e.g., A, B, C) and lower-case letters in italics denote their values (e.g., a, b, c). Similarly, upper-case bold letters denote random variable sets (e.g., $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$) and lower-case bold letters denote their values (e.g., $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$). The terms "variables", "genes", and "vertices" are used interchangeably in this work. If a graph contains an edge $X \rightarrow Y$, then X is a *parent* of Y and Y is a *child* of X. A vertex X is a *spouse* of Y if they share a common child. An undirected edge $X - Y$ denotes *adjacency relation* between X and Y (i.e., presence of an edge directly connecting X and Y). A *path* p is a set of consecutive edges (independent of the direction) without visiting a vertex more than once. A *directed path* p from X to Y is a set of consecutive edges with direction "→" connecting X with Y, i.e. $X \rightarrow \ldots \rightarrow Y$. X is an *ancestor* of Y (and at the same time Y is a *descendant* of X) if there exists a directed path p from X to Y. A *directed cycle* is a nonempty directed path that starts and ends on the same vertex X. Four classes of graphs are considered in this work:

- *Directed graphs*: Directed graphs where vertices can be connected only with an edge "→".
- *Directed acyclic graphs* (DAGs): Directed graphs without directed cycles where vertices can be connected only with an edge "→".
- *Ancestral graphs*: Directed graphs[7] without directed cycles where vertices can be connected with one of the two edges: "→" or "↔". For any two vertices X and Y, if there is an edge $X \leftrightarrow Y$, then X is not an ancestor of Y and Y is not an ancestor of X. In other

[7] Notice that I follow (Zhang and Spirtes, 2005) and consider only *directed* ancestral graphs.

words, X and Y have a hidden confounder (Zhang and Spirtes, 2005; Richardson and Spirtes, 2002).

- *Maximal ancestral graphs* (MAGs): Ancestral graphs with the following property: for any two non-adjacent vertices there is a set of vertices that m-separates them (the definition of m-separation is given below) (Zhang and Spirtes, 2005; Richardson and Spirtes, 2002).

Definition of <u>conditional independence</u>: Two sets of variables **X** and **Y** are conditionally independent given a set of variables **Z** in the joint probability distribution $\mathbb{P}$ (denoted as $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$) if P(**X**=**x** | **Y**=**y**, **Z**=**z**) = P(**X**=**x** | **Z**=**z**) whenever P(**Y**=**y**, **Z**=**z**) > 0.

For notational convenience conditional dependence is defined as absence of conditional independence and denoted as $\mathbf{X} \not\perp \mathbf{Y} \mid \mathbf{Z}$. When two sets of variables **X** and **Y** are conditionally independent given an empty set, I simply say that they are independent and denote this by $\mathbf{X} \perp \mathbf{Y}$. Similarly the dependence of **X** and **Y** is defined and denoted as $\mathbf{X} \not\perp \mathbf{Y}$.

Definition of <u>collider</u>: A vertex W on the path p is a collider if p contains two incoming edges into W (i.e., $\exists X$ and Y: $X \rightarrow W \leftarrow Y$, or $X \leftrightarrow W \leftarrow Y$, or $X \rightarrow W \leftrightarrow Y$, or $X \leftrightarrow W \leftrightarrow Y$ and $\{X, W, Y\} \subseteq p$).

Definition of <u>blocked path</u>: A path p from X to Y is blocked by a set of vertices **Z** if there is a vertex W on the path p for which one of the two conditions hold: (i) W is not a collider and $W \in \mathbf{Z}$, or (ii) W is a collider and neither W nor its descendants are in **Z**.

Definition of <u>d-separation</u>: X is d-separated from Y given **Z** in directed graph $\mathbb{G}$ if every path in $\mathbb{G}$ from X to Y is blocked by **Z**.

Definition of <u>m-separation</u>: X is m-separated from Y given **Z** in ancestral graph $\mathbb{G}$ if every path in $\mathbb{G}$ from X to Y is blocked by **Z**.

Definition of <u>local Markov condition</u>: The joint probability distribution $\mathbb{P}$ over variables **V** satisfies the local Markov condition for a directed acyclic graph (DAG) $\mathbb{G} = <\mathbf{V}, \mathbb{E}>$ if and only

if for each *W* in **V**, *W* is independent of all variables in **V** excluding descendants of *W* and parents of *W* given parents of *W* (Richardson and Spirtes, 1999).

The definition below extends Markov condition to any directed and ancestral graphs, not necessarily DAGs:

Definition of global Markov condition: The joint probability distribution $\mathbb{P}$ over variables **V** satisfies the global Markov condition for a directed graph (ancestral graph) $\mathbb{G}$ = <**V**, $\mathbb{E}$> if and only if for any three disjoint subsets of variables **X**, **Y**, **Z** from **V**, if **X** is d-separated (m-separated) from **Y** given **Z** in $\mathbb{G}$ then **X** is independent of **Y** given **Z** in $\mathbb{P}$ (Richardson and Spirtes, 2002; Richardson and Spirtes, 1999).

It follows that if the underlying graph $\mathbb{G}$ is a DAG, then the global Markov condition is equivalent to the local Markov condition (Richardson and Spirtes, 1999).

Definition of Bayesian network: $\mathbb{N}$ = <$\mathbb{G}$, $\mathbb{P}$> is a Bayesian network if $\mathbb{P}$ satisfies the local Markov condition for a DAG $\mathbb{G}$.

APPENDIX B

FAITHFULNESS ASSUMPTION AND EXTENSIONS

Definition of DAG-faithfulness: If all and only conditional independence relations true in $\mathbb{P}$ defined over variables **V** are entailed by the local Markov condition applied to a DAG $\mathbb{G}$ = <**V**, $\mathbb{E}$>, then $\mathbb{P}$ and $\mathbb{G}$ are DAG-faithful to one another (Spirtes et al., 2000).

The definition below extends faithfulness to any directed or ancestral graphs, not necessarily DAGs:

Definition of graph-faithfulness: If all and only conditional independence relations true in $\mathbb{P}$ defined over variables **V** are entailed by the global Markov condition applied to a directed or ancestral graph $\mathbb{G}$ = <**V**, $\mathbb{E}$>, then $\mathbb{P}$ and $\mathbb{G}$ are graph-faithful to one another.

Alternatively, $\mathbb{P}$ and $\mathbb{G}$ are DAG-faithful to one another if the following two conditions hold (Neapolitan, 2004): (i) $\mathbb{P}$ satisfies the local Markov condition for $\mathbb{G}$; and (ii) the only conditional independencies in $\mathbb{P}$ are those entailed by the local Markov condition for $\mathbb{G}$. Similarly, $\mathbb{P}$ and $\mathbb{G}$ are graph-faithful to one another if: (i) $\mathbb{P}$ satisfies the global Markov condition for $\mathbb{G}$; and (ii) the only conditional independencies in $\mathbb{P}$ are those entailed by the global Markov condition for $\mathbb{G}$. It follows that if $\mathbb{G}$ is a DAG, then DAG-faithfulness and graph-faithfulness are equivalent.

A relaxed version of the faithfulness assumption is given below:

Definition of adjacency faithfulness: Given a directed or ancestral graph $\mathbb{G}$ = <**V**, $\mathbb{E}$> and a joint probability distribution $\mathbb{P}$ defined over variables **V**, $\mathbb{P}$ and $\mathbb{G}$ are adjacency faithful to one another if every adjacency relation between X and Y in $\mathbb{G}$ implies that X and Y are conditionally dependent given any subset of **V** \ {X, Y} in $\mathbb{P}$ (Ramsey et al., 2006).

Consider the following example given in (Ramsey et al., 2006). A Bayesian network is specified by the graph $A \rightarrow B \rightarrow C$ and the joint probability distribution where only two independence relations hold: $A \perp C \mid B$ and $A \perp C$. Clearly, this graph is not DAG-faithful (or graph-faithful) to the joint probability distribution because the independence relation $A \perp C$ is not entailed by the local (or global) Markov condition. On the other hand, the adjacency faithfulness is not violated in this example. Also, notice that unlike DAG-faithfulness or graph-faithfulness, adjacency faithfulness does not imply that the Markov condition holds.

The adjacency faithfulness assumption can be further relaxed to focus on the specific response variable of interest:

Definition of <u>local adjacency faithfulness with respect to a variable:</u> Given a directed or ancestral graph $\mathbb{G}$ = <**V**, $\mathbb{E}$> and a joint probability distribution $\mathbb{P}$ defined over variables **V**, $\mathbb{P}$ and $\mathbb{G}$ are locally adjacency faithful with respect to T if every adjacency relation between T and X in $\mathbb{G}$ implies that T and X are conditionally dependent given any subset of **V** \ $\{T, X\}$ in $\mathbb{P}$.

Next, I introduce another relaxed version of faithfulness:

Definition of <u>path faithfulness:</u> Given a directed or ancestral graph $\mathbb{G}$ = <**V**, $\mathbb{E}$> and a joint probability distribution $\mathbb{P}$ defined over variables **V**, $\mathbb{P}$ and $\mathbb{G}$ are path faithful to one another if for every path p without colliders: $Y - X_1 - X_2 - \ldots - X_M$, the following condition holds for every $k = 1,\ldots, M$: Y and X_k are conditionally dependent given any subset of **V** \ $\{Y, X_1, \ldots, X_k\}$ in $\mathbb{P}$.

The above definition does not imply that the Markov condition holds. Also notice that if $\mathbb{P}$ and $\mathbb{G}$ are path faithful to one another, then they are also adjacency faithful to one another. However, the converse may not be true in general.

The path faithfulness assumption is further relaxed to focus on the specific response variable of interest:

Definition of <u>local path faithfulness with respect to a variable:</u> Given a directed or ancestral graph $\mathbb{G} = <\mathbf{V}, \mathbb{E}>$ and a joint probability distribution $\mathbb{P}$ defined over variables $\mathbf{V}$, $\mathbb{P}$ and $\mathbb{G}$ are locally path faithful with respect to T if for every path p without colliders that involves a variable T: $T - X_1 - X_2 - \ldots - X_M$, the following condition holds for every $k = 1,\ldots, M$: T and X_k are conditionally dependent given any subset of $\mathbf{V} \setminus \{T, X_1, \ldots, X_k\}$ in $\mathbb{P}$.

APPENDIX C

REVISED PROOFS OF CORRECTNESS FOR TWO MARKOV BOUNDARY ALGORITHMS

Theorem 6: IAMB outputs a Markov boundary of *T* if the joint probability distribution $\mathbb{P}$ satisfies the local composition property with respect to *T*.

Proof: First I prove that **M** is a Markov blanket of *T* at the end of Phase I. Suppose it is not, i.e. $T \not\perp (\mathbf{V} \setminus \mathbf{M} \setminus \{T\}) \mid \mathbf{M}$. By the local composition property with respect to *T*, there exists $X \in (\mathbf{V} \setminus \mathbf{M} \setminus \{T\})$ such that $T \not\perp X \mid \mathbf{M}$. This contradicts the exit condition from the loop in line 6 that states that **M** should not change in the present iteration which can be the case if and only if for every $X \in (\mathbf{V} \setminus \mathbf{M} \setminus \{T\})$, $T \perp X \mid \mathbf{M}$. Therefore, **M** is a Markov blanket of *T* at the end of Phase I.

Next I prove that **M** remains a Markov blanket of *T* at the end of Phase II. Assume that a variable $X \in \mathbf{M}$ can be rendered independent from *T* by conditioning on the remaining variables in **M**, i.e. $T \perp X \mid (\mathbf{M} \setminus \{X\})$. From Phase I it follows that $T \perp (\mathbf{V} \setminus \mathbf{M} \setminus \{T\}) \mid \mathbf{M}$. The above two independence relations by the contraction property imply that $T \perp (\mathbf{V} \setminus (\mathbf{M} \setminus \{X\}) \setminus \{T\}) \mid (\mathbf{M} \setminus \{X\})$. Thus, **M** is a Markov blanket of *T* at the end of Phase II of the algorithm.

Finally I prove that **M** is a Markov boundary of *T* at the end of Phase II. Suppose it is not and thus there exists $\mathbf{N} \subset \mathbf{M}$ that is a Markov blanket of *T*. Let $X \in \mathbf{M} \setminus \mathbf{N}$ and $\mathbf{Y} \subseteq (\mathbf{V} \setminus \mathbf{N} \setminus \{T\} \setminus \{X\})$. By definition of the Markov blanket, $T \perp (\mathbf{V} \setminus \mathbf{N} \setminus \{T\}) \mid \mathbf{N}$. By the decomposition property, $T \perp (\mathbf{Y} \cup \{X\}) \mid \mathbf{N}$. The latter independence relation implies $T \perp X \mid (\mathbf{N} \cup \mathbf{Y})$ by the weak union property. Therefore, any variable $X \in \mathbf{M} \setminus \mathbf{N}$ would be

removed by the algorithm in line 9 which contradicts the assumption that the algorithm output **M** and $\mathbf{N} \subset \mathbf{M}$ is another Markov blanket of *T*. Therefore, **M** is a Markov boundary of *T* at the end of Phase II. (Q.E.D.)

Theorem 8: HITON-PC outputs a Markov boundary of *T* if (i) the joint probability distribution $\mathbb{P}$ and directed or ancestral graph $\mathbb{G}$ are locally adjacency faithful with respect to *T* with the exception of violations of the intersection property; (ii) $\mathbb{P}$ satisfies the global Markov condition for $\mathbb{G}$; (iii) the set of vertices adjacent with *T* in $\mathbb{G}$ is a Markov blanket of *T*.

Proof: First I prove that the set **M** is a Markov blanket of *T* in line 12 of the algorithm. Assumptions (i) and (iii) imply that all Markov blanket members will be in the set **E** after line 4. Notice that violations of the intersection property do not affect the above statement. In lines 8 and 11, *X* can be removed from **M** because it is either a non-Markov boundary member or the intersection property is violated. The former case does not compromise the Markov blanket property of **M**, thus I consider only the latter case. Since the intersection property is violated, the following relations hold in $\mathbb{P}$: $T \perp X \,|\, \mathbf{Z}$, $T \perp \mathbf{Z} \,|\, X$ and $T \not\perp (\{X\} \cup \mathbf{Z})$. Below I show that if *X* is a member of some Markov blanket $\mathbf{M}_1 = \mathbf{N} \cup \{X\}$, then $\mathbf{M}_2 = \mathbf{N} \cup \mathbf{Z}$ is also a Markov blanket where $X \notin \mathbf{N}$ and $\mathbf{Z} \cap \mathbf{N} = \emptyset$. Since $\mathbf{M}_1$ is a Markov blanket, $T \perp (\mathbf{V} \setminus \{T\} \setminus (\mathbf{N} \cup \{X\})) \,|\, (\mathbf{N} \cup \{X\})$. By the self-conditioning property, it follows that $T \perp (\mathbf{V} \setminus \{T\}) \,|\, (\mathbf{N} \cup \{X\})$. The previous independence relation is equivalent to $T \perp ((\mathbf{V} \setminus \{T\} \setminus \mathbf{Z}) \cup \mathbf{Z}) \,|\, (\mathbf{N} \cup \{X\})$. By the weak union property, it follows that $T \perp (\mathbf{V} \setminus \{T\} \setminus \mathbf{Z}) \,|\, (\mathbf{N} \cup \{X\} \cup \mathbf{Z})$. By the self-conditioning properly, it follows that $T \perp (\mathbf{V} \setminus \{T\}) \,|\, (\mathbf{N} \cup \{X\} \cup \mathbf{Z})$. Equivalently, $T \perp (\mathbf{V} \setminus \{T\}) \,|\, ((\mathbf{N} \cup \{X\}) \cup (\mathbf{N} \cup \mathbf{Z}))$. Since, $T \perp X \,|\, \mathbf{Z}$, by the self-conditioning property $T \perp (\mathbf{N} \cup \{X\}) \,|\, (\mathbf{N} \cup \mathbf{Z})$. By the contraction property, $T \perp (\mathbf{V} \setminus \{T\}) \,|\, ((\mathbf{N} \cup \{X\}) \cup (\mathbf{N} \cup \mathbf{Z}))$ and $T \perp (\mathbf{N} \cup \{X\}) \,|\, (\mathbf{N} \cup \mathbf{Z})$ imply that $T \perp ((\mathbf{V} \setminus \{T\}) \cup (\mathbf{N} \cup \{X\})) \,|\, (\mathbf{N} \cup \mathbf{Z})$. This is equivalent to $T \perp (\mathbf{V} \setminus \{T\}) \,|\, (\mathbf{N} \cup \mathbf{Z})$. By the

decomposition property this implies that $\mathbf{M}_2 = \mathbf{N} \cup \mathbf{Z}$ is a Markov blanket of *T*. Therefore the set **M** is a Markov blanket of *T* after line 12 of the algorithm.

Now I prove that the set **M** retuned by HITON-PC is a Markov boundary of *T*. Suppose it is not and thus there exists $\mathbf{N} \subset \mathbf{M}$ that is a Markov blanket of *T*. Let $X \in \mathbf{M} \setminus \mathbf{N}$ and $\mathbf{Y} \subseteq (\mathbf{V} \setminus \mathbf{N} \setminus \{T\} \setminus \{X\})$. By the definition of Markov blanket, $T \perp (\mathbf{V} \setminus \mathbf{N} \setminus \{T\}) \mid \mathbf{N}$. By the decomposition property, $T \perp (\mathbf{Y} \cup \{X\}) \mid \mathbf{N}$. The latter independence relation implies $T \perp X \mid (\mathbf{N} \cup \mathbf{Y})$ by the weak union property. Therefore, any variable $X \in \mathbf{M} \setminus \mathbf{N}$ would be removed by the algorithm in line 11 which contradicts the assumption that the algorithm output **M** and $\mathbf{N} \subset \mathbf{M}$ is another Markov blanket of *T*. Therefore, HITON-PC outputs a Markov boundary of *T*. (Q.E.D.)

The proofs of correctness provided above for Markov boundary algorithms implicitly assume that the base statistical decisions about dependence and independence are correct. This requirement is satisfied when the dataset $\mathbb{D}$ is a large i.i.d. (independent and identically distributed) sample of the underlying probability distribution $\mathbb{P}$. When the sample size is small, the statistical test of null hypothesis of independence will incur type I and II errors. This may affect correctness of the algorithm's output Markov boundary.

APPENDIX D

TIED1 AND *TIED2* NETWORK STRUCTURE AND PARAMETERIZATION

Using the principles from Figure 1 of the thesis, a discrete artificial network *TIED1* with 30 variables (including a response variable T) was constructed. Figure 17 shows the network structure and specifies which variables contain the same information about T by the color of highlighting. For example, variables X_{12}, X_{13}, and X_{14} provide exactly the same information about T and thus are interchangeable for prediction of T. The parameterization of the network is provided in Table 18. The network contains 72 Markov boundaries of T. Each of these Markov boundaries contains 5 variables: (i) X_{10}, (ii) X_5 or X_9, (iii) X_{12} or X_{13} or X_{14}, (iv) X_{19} or X_{20} or X_{21}, and (v) X_1 or X_2 or X_3 or X_{11}.

A discrete artificial network *TIED2* with 1,000 variables (including a response variable T) was constructed by augmenting *TIED1* network with a total of 970 variables such that the resulting network has exactly the same 72 Markov boundaries. Out of 970 variables that were added to the prior network, 110 variables have a path to T and 860 variables do not.

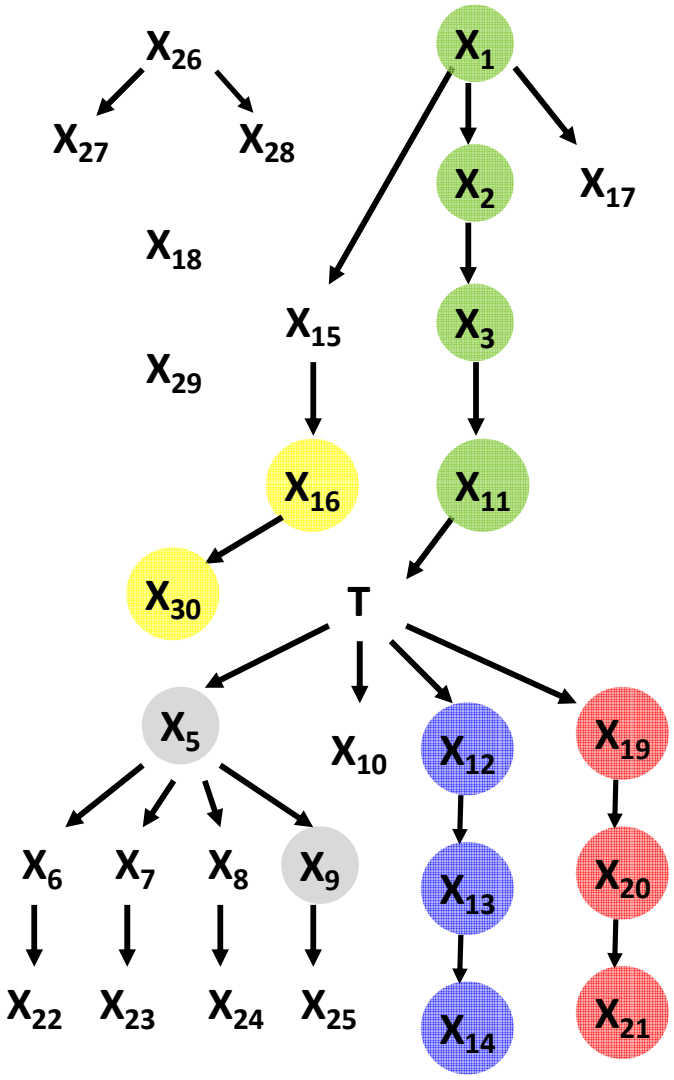


Figure 17: Graphical visualization of a discrete artificial network *TIED1* with 30 variables (including a response variable *T*). Variables that contain exactly the same information about *T* are highlighted with the same color, e.g. variables X_{12}, X_{13}, and X_{14} provide exactly the same information about *T* and thus are interchangeable for prediction of *T*.

X_1: P(X_1=0) = 0.25 P(X_1=1) = 0.25 P(X_1=2) = 0.25 P(X_1=3) = 0.25	X_6: P(X_6=0|X_5=0) = 0.6 P(X_6=1|X_5=0) = 0.2 P(X_6=2|X_5=0) = 0.2 P(X_6=0|X_5=1) = 0.5 P(X_6=1|X_5=1) = 0.25 P(X_6=2|X_5=1) = 0.25 P(X_6=0|X_5=2) = 0.8 P(X_6=1|X_5=2) = 0.1 P(X_6=2|X_5=2) = 0.1	X_{11}: P(X_{11}=0|X_3=0) = 1.0 P(X_{11}=0|X_3=1) = 1.0 P(X_{11}=1|X_3=2) = 0.3 P(X_{11}=2|X_3=2) = 0.7 P(X_{11}=3|X_3=3) = 1.0
X_2: P(X_2=0|X_1=0) = 0.8 P(X_2=1|X_1=0) = 0.2 P(X_2=0|X_1=1) = 0.1 P(X_2=1|X_1=1) = 0.9 P(X_2=2|X_1=2) = 1.0 P(X_2=3|X_1=3) = 1.0	X_7: P(X_7=1|X_5=0) = 0.5 P(X_7=2|X_5=0) = 0.5 P(X_7=0|X_5=1) = 0.8 P(X_7=1|X_5=1) = 0.2 P(X_7=0|X_5=2) = 0.2 P(X_7=1|X_5=2) = 0.3 P(X_7=2|X_5=2) = 0.5	X_{12}: P(X_{12}=0|T=0) = 1.0 P(X_{12}=0|T=1) = 1.0 P(X_{12}=0|T=2) = 1.0 P(X_{12}=1|T=3) = 0.5 P(X_{12}=2|T=3) = 0.5
X_3: P(X_3=0|X_2=0) = 0.3 P(X_3=1|X_2=0) = 0.7 P(X_3=0|X_2=1) = 0.8 P(X_3=1|X_2=1) = 0.2 P(X_3=2|X_2=2) = 1.0 P(X_3=3|X_2=3) = 1.0	X_8: P(X_8=0|X_5=0) = 0.9 P(X_8=1|X_5=0) = 0.1 P(X_8=0|X_5=1) = 0.7 P(X_8=1|X_5=1) = 0.2 P(X_8=2|X_5=1) = 0.1 P(X_8=0|X_5=2) = 0.6 P(X_8=1|X_5=2) = 0.3 P(X_8=2|X_5=2) = 0.1	X_{13}: P(X_{13}=0|X_{12}=0) = 1.0 P(X_{13}=1|X_{12}=1) = 0.5 P(X_{13}=2|X_{12}=1) = 0.5 P(X_{13}=1|X_{12}=2) = 0.5 P(X_{13}=2|X_{12}=2) = 0.5
T: P(T=0|X_{11}=0) = 1.0 P(T=0|X_{11}=1) = 1.0 P(T=0|X_{11}=2) = 1.0 P(T=1|X_{11}=3) = 0.3 P(T=2|X_{11}=3) = 0.3 P(T=3|X_{11}=3) = 0.4	X_9: P(X_9=1|X_5=0) = 1.0 P(X_9=2|X_5=1) = 1.0 P(X_9=0|X_5=2) = 1.0	X_{14}: P(X_{14}=0|X_{13}=0) = 1.0 P(X_{14}=1|X_{13}=1) = 0.5 P(X_{14}=2|X_{13}=1) = 0.5 P(X_{14}=1|X_{13}=2) = 0.5 P(X_{14}=2|X_{13}=2) = 0.5
X_5: P(X_5=1|T=0) = 0.9 P(X_5=2|T=0) = 0.1 P(X_5=0|T=1) = 0.8 P(X_5=1|T=1) = 0.1 P(X_5=2|T=1) = 0.1 P(X_5=0|T=2) = 0.1 P(X_5=1|T=2) = 0.8 P(X_5=2|T=2) = 0.1 P(X_5=0|T=3) = 0.1 P(X_5=1|T=3) = 0.1 P(X_5=2|T=3) = 0.8	X_{10}: P(X_{10}=0|T=0) = 0.1 P(X_{10}=1|T=0) = 0.8 P(X_{10}=2|T=0) = 0.1 P(X_{10}=1|T=1) = 0.1 P(X_{10}=2|T=1) = 0.9 P(X_{10}=0|T=2) = 0.1 P(X_{10}=1|T=2) = 0.8 P(X_{10}=2|T=2) = 0.1 P(X_{10}=0|T=3) = 0.2 P(X_{10}=1|T=3) = 0.7 P(X_{10}=2|T=3) = 0.1	X_{15}: P(X_{15}=0|X_1=0) = 0.8 P(X_{15}=1|X_1=0) = 0.1 P(X_{15}=2|X_1=0) = 0.1 P(X_{15}=0|X_1=1) = 0.1 P(X_{15}=1|X_1=1) = 0.8 P(X_{15}=2|X_1=1) = 0.1 P(X_{15}=0|X_1=2) = 0.8 P(X_{15}=1|X_1=2) = 0.1 P(X_{15}=2|X_1=2) = 0.1 P(X_{15}=0|X_1=3) = 0.1 P(X_{15}=1|X_1=3) = 0.1 P(X_{15}=2|X_1=3) = 0.8

Table 18 (continued on the next page): Parameterization of the *TIED1* network. Only nonzero probabilities are shown in the table.

X_{16}: P(X_{16}=0|X_{15}=0) = 1.0 P(X_{16}=0|X_{15}=1) = 1.0 P(X_{16}=1|X_{15}=2) = 0.5 P(X_{16}=2|X_{15}=2) = 0.5	X_{21}: P(X_{21}=0|X_{20}=0) = 1.0 P(X_{21}=1|X_{20}=1) = 1.0 P(X_{21}=2|X_{20}=2) = 1.0	X_{26}: P(X_{26}=0) = 0.5 P(X_{26}=1) = 0.5
X_{17}: P(X_{17}=0|X_{1}=0) = 0.2 P(X_{17}=1|X_{1}=0) = 0.6 P(X_{17}=2|X_{1}=0) = 0.2 P(X_{17}=0|X_{1}=1) = 0.1 P(X_{17}=1|X_{1}=1) = 0.3 P(X_{17}=2|X_{1}=1) = 0.6 P(X_{17}=0|X_{1}=2) = 0.5 P(X_{17}=1|X_{1}=2) = 0.1 P(X_{17}=2|X_{1}=2) = 0.4 P(X_{17}=0|X_{1}=3) = 0.3 P(X_{17}=1|X_{1}=3) = 0.5 P(X_{17}=2|X_{1}=3) = 0.2	X_{22}: P(X_{22}=0|X_{6}=0) = 0.2 P(X_{22}=1|X_{6}=0) = 0.6 P(X_{22}=2|X_{6}=0) = 0.2 P(X_{22}=0|X_{6}=1) = 0.1 P(X_{22}=1|X_{6}=1) = 0.3 P(X_{22}=2|X_{6}=1) = 0.6 P(X_{22}=0|X_{6}=2) = 0.5 P(X_{22}=1|X_{6}=2) = 0.1 P(X_{22}=2|X_{6}=2) = 0.4	X_{27}: P(X_{27}=0|X_{26}=0) = 0.1 P(X_{27}=1|X_{26}=0) = 0.9 P(X_{27}=0|X_{26}=1) = 0.3 P(X_{27}=1|X_{26}=1) = 0.7
X_{18}: P(X_{18}=0) = 0.25 P(X_{18}=1) = 0.25 P(X_{18}=2) = 0.25 P(X_{18}=3) = 0.25	X_{23}: P(X_{23}=0|X_{7}=0) = 0.3 P(X_{23}=1|X_{7}=0) = 0.2 P(X_{23}=2|X_{7}=0) = 0.5 P(X_{23}=0|X_{7}=1) = 0.8 P(X_{23}=1|X_{7}=1) = 0.1 P(X_{23}=2|X_{7}=1) = 0.1 P(X_{23}=0|X_{7}=2) = 0.6 P(X_{23}=1|X_{7}=2) = 0.2 P(X_{23}=2|X_{7}=2) = 0.2	X_{28}: P(X_{28}=0|X_{26}=0) = 0.4 P(X_{28}=1|X_{26}=0) = 0.6 P(X_{28}=0|X_{26}=1) = 0.8 P(X_{28}=1|X_{26}=1) = 0.2
X_{19}: P(X_{19}=1|T=0) = 0.1 P(X_{19}=2|T=0) = 0.9 P(X_{19}=0|T=1) = 0.1 P(X_{19}=2|T=1) = 0.9 P(X_{19}=0|T=2) = 0.8 P(X_{19}=1|T=2) = 0.1 P(X_{19}=2|T=2) = 0.1 P(X_{19}=0|T=3) = 0.1 P(X_{19}=1|T=3) = 0.8 P(X_{19}=2|T=3) = 0.1	X_{24}: P(X_{24}=0|X_{8}=0) = 0.5 P(X_{24}=1|X_{8}=0) = 0.1 P(X_{24}=2|X_{8}=0) = 0.4 P(X_{24}=0|X_{8}=1) = 0.6 P(X_{24}=1|X_{8}=1) = 0.3 P(X_{24}=2|X_{8}=1) = 0.1 P(X_{24}=0|X_{8}=2) = 0.7 P(X_{24}=1|X_{8}=2) = 0.1 P(X_{24}=2|X_{8}=2) = 0.2	X_{29}: P(X_{29}=0) = 0.33 P(X_{29}=1) = 0.33 P(X_{29}=2) = 0.33
X_{20}: P(X_{20}=1|X_{19}=0) = 1.0 P(X_{20}=2|X_{19}=1) = 1.0 P(X_{20}=0|X_{19}=2) = 1.0	X_{25}: P(X_{25}=0|X_{9}=0) = 0.8 P(X_{25}=1|X_{9}=0) = 0.1 P(X_{25}=2|X_{9}=0) = 0.1 P(X_{25}=0|X_{9}=1) = 0.6 P(X_{25}=1|X_{9}=1) = 0.2 P(X_{25}=2|X_{9}=1) = 0.2 P(X_{25}=0|X_{9}=2) = 0.5 P(X_{25}=1|X_{9}=2) = 0.3 P(X_{25}=2|X_{9}=2) = 0.2	X_{30}: P(X_{30}=0|X_{16}=0) = 1.0 P(X_{30}=1|X_{16}=1) = 0.5 P(X_{30}=2|X_{16}=1) = 0.5 P(X_{30}=1|X_{16}=2) = 0.5 P(X_{30}=2|X_{16}=2) = 0.5

Table 18 (continued from the previous page)

APPENDIX E

LIND NETWORK STRUCTURE AND PARAMETERIZATION

Figure 18 shows the network structure and specifies which variables contain the same information about T by the color of highlighting. Table 19 provides details about parameterization. For example, variables X_8, X_3, and X_{17} provide exactly the same information about T and thus are interchangeable for prediction of T. Similarly, variable X_7 and a variable set $\{X_1, X_2\}$ provide the same information about T. The network contains 12 Markov boundaries of T. Each of these Markov boundaries contains 3 or 5 variables: (i) X_7 or $\{X_1, X_2\}$, (ii) X_8 or X_3 or X_{17}, and (iii) X_9 or $\{X_4, X_5\}$.

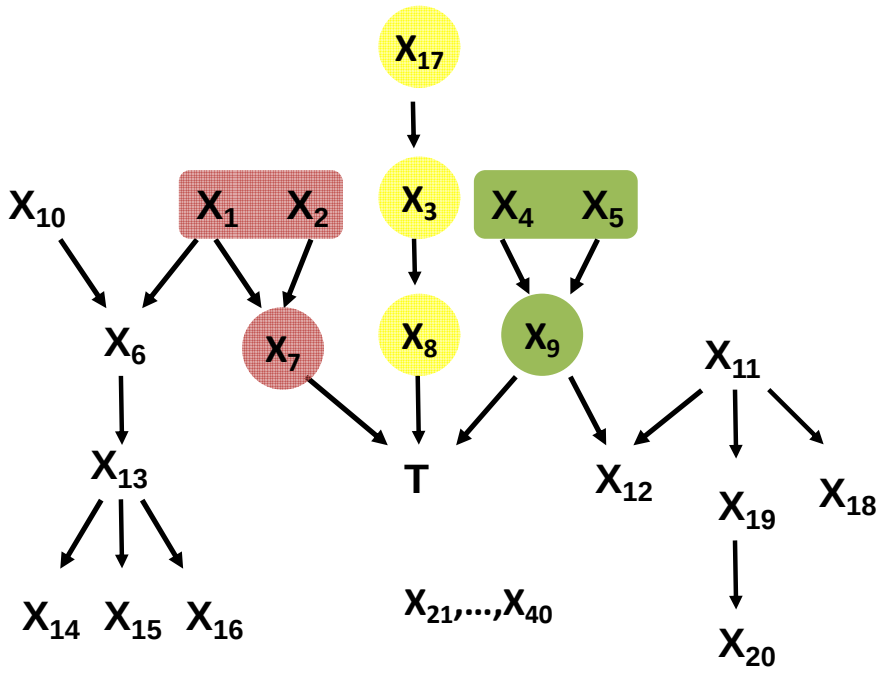


Figure 18: Graphical visualization of a continuous artificial network *LIND* with 41 variables (including a response variable T). Variables that contain exactly the same information about T are highlighted with the same color, e.g. variables X_8, X_3, and X_{17} provide exactly the same information about T and thus are interchangeable for prediction of T. Similarly, variable X_7 and a variable set $\{X_1, X_2\}$ provide the same information about T.

$X_1 = \mathbb{N}(0,1)$	$X_8 = 0.9X_3$	$X_{15} = 0.7X_{13} + 0.2\mathbb{N}(0,1)$
$X_2 = \mathbb{N}(0,1)$	$X_9 = 0.9X_4 + 0.7X_5$	$X_{16} = 0.9X_{13} + 0.2\mathbb{N}(0,1)$
$X_3 = 0.9X_{17}$	$X_{10} = \mathbb{N}(0,1)$	$X_{17} = \mathbb{N}(0,1)$
$X_4 = \mathbb{N}(0,1)$	$X_{11} = \mathbb{N}(0,1)$	$X_{18} = 0.6X_{11} + 0.2\mathbb{N}(0,1)$
$X_5 = \mathbb{N}(0,1)$	$X_{12} = 0.7X_{11} + 0.3X_9 + 0.1\mathbb{N}(0,1)$	$X_{19} = 0.9X_{11} + 0.1\mathbb{N}(0,1)$
$X_6 = 0.8X_{10} + 0.4X_1 + 0.1\mathbb{N}(0,1)$	$X_{13} = 0.7X_6 + 0.1\mathbb{N}(0,1)$	$X_{20} = 0.8X_{19} + 0.1\mathbb{N}(0,1)$
$X_7 = 0.7X_1 + 0.8X_2$	$X_{14} = 0.8X_{13} + 0.1\mathbb{N}(0,1)$	$X_{21},\ldots,X_{40} = \mathbb{N}(0,1)$
$T = (0.8X_7 + 0.9X_8 + 0.8X_9 + 0.2\mathbb{N}(0,1)) > 0$		

Table 19: Parameterization of the *LIND* network. $\mathbb{N}(0,1)$ denotes a random Normal variable with mean = 0 and standard deviation = 1.

APPENDIX F

XORD NETWORK STRUCTURE AND PARAMETERIZATION

Figure 19 shows the network structure and specifies which variables contain the same information about T by the color of highlighting. Table 20 provides details about parameterization. For example, variables X_1 and X_5 provide exactly the same information about T and thus are interchangeable for prediction of T. Similarly, variable X_9 and each of the four variable sets $\{X_5, X_6\}$, $\{X_1, X_2\}$, $\{X_1, X_6\}$, $\{X_5, X_2\}$ provide the same information about T. The network contains 25 Markov boundaries of T. Each of these Markov boundaries contains 3 or 5 variables: (i) X_9 or $\{X_5, X_6\}$ or $\{X_1, X_2\}$ or $\{X_1, X_6\}$ or $\{X_5, X_2\}$, (ii) X_{10}, and (iii) X_{11} or $\{X_7, X_8\}$ or $\{X_3, X_4\}$ or $\{X_3, X_8\}$ or $\{X_7, X_4\}$.

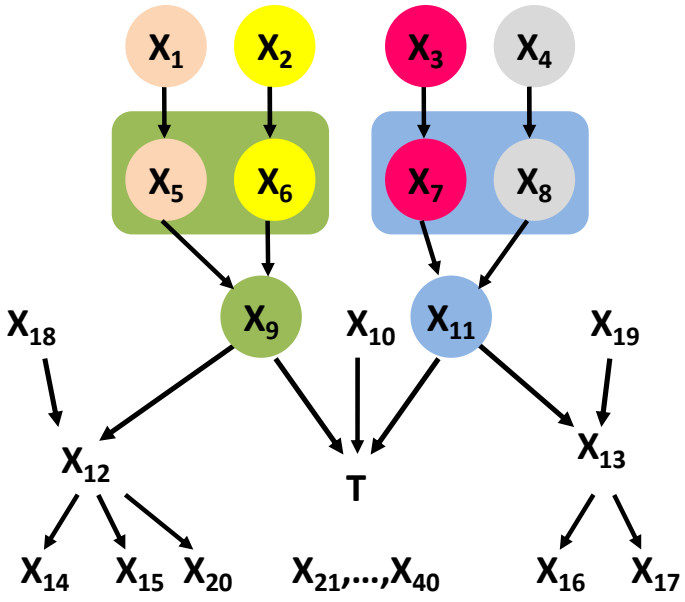


Figure 19: Graphical visualization of a discrete artificial network *XORD* with 41 variables (including a response variable T). All variables take binary values {0, 1}. Variables that contain exactly the same information about T are highlighted with the same color, e.g. variables X_1 and X_5 provide exactly the same information about T and thus are interchangeable for prediction of T. Similarly, variable X_9 and each of the four variable sets $\{X_5, X_6\}$, $\{X_1, X_2\}$, $\{X_1, X_6\}$, $\{X_5, X_2\}$ provide the same information about T.

X_1: P(X_1=0) = 0.5	$X_8 = X_4$	X_{15}: P(X_{15}=0|X_{12}=0) = 0.3 P(X_{15}=0|X_{12}=1) = 0.1
X_2: P(X_2=0) = 0.5	X_9 = OR(X_5, X_6)	X_{16}: P(X_{16}=0|X_{13}=0) = 0.2 P(X_{16}=0|X_{13}=1) = 0.5
X_3: P(X_3=0) = 0.5	X_{10}: P(X_{10}=0) = 0.5	X_{17}: P(X_{17}=0|X_{13}=0) = 0.6 P(X_{17}=0|X_{13}=1) = 0.4
X_4: P(X_4=0) = 0.5	X_{11} = OR(X_7, X_8)	X_{18}: P(X_{18}=0) = 0.5
$X_5 = 1 - X_1$	X_{12}: P(X_{12}=0|X_{18}=0, X_9=0) = 0.4 P(X_{12}=0|X_{18}=0, X_9=1) = 0.5 P(X_{12}=0|X_{18}=1, X_9=0) = 0.5 P(X_{12}=0|X_{18}=1, X_9=1) = 0.6	X_{19}: P(X_{18}=0) = 0.5
$X_6 = X_2$	X_{13}: P(X_{13}=0|X_{11}=0, X_{19}=0) = 0.4 P(X_{13}=0|X_{11}=0, X_{19}=1) = 0.6 P(X_{13}=0|X_{11}=1, X_{19}=0) = 0.5 P(X_{13}=0|X_{11}=1, X_{19}=1) = 0.5	X_{20}: P(X_{20}=0|X_{12}=0) = 0.5 P(X_{20}=0|X_{12}=1) = 0.2
$X_7 = 1 - X_3$	X_{14}: P(X_{14}=0|X_{12}=0) = 0.2 P(X_{14}=0|X_{12}=1) = 0.4	X_i: P(X_i=0) = 0.5, i = 21,…,40.
T = XOR(X_9, X_{10}, X_{11})		

Table 20: Parameterization of the *XORD* network. OR and XOR denote corresponding binary functions.

APPENDIX G

STATE-OF-THE-ART ALGORITHMS FOR MULTIPLE SIGNATURE IDENTIFICATION USED IN COMPUTATIONAL EXPERIMENTS

Eight state-of-the-art methods to extract multiple signatures and compare to TIE* were used in experiments. These algorithms were executed on Intel Xeon 2.4 GHz CPUs for up to one week of single-CPU time or to produce up to 5,000 signatures (per method and dataset), whichever termination criterion was met first.

Four methods were resampling-based techniques that apply a signature extraction algorithm to bootstrap samples of the original dataset. The following signature extraction algorithms were used: (i) SVM-based recursive feature elimination (SVM-RFE) (Guyon et al., 2002); (ii) SVM-RFE with additional application of a formal statistical comparison test[8] to identify the most parsimonious signature with predictivity statistically indistinguishable from the observed best one; (iii) backward wrapping based on univariate ranking of variables by Kruskal-Wallis non-parametric ANOVA (Statnikov et al., 2005; Hollander and Wolfe, 1999); and (iv) backward wrapping based on Kruskal-Wallis ANOVA with additional statistical comparison step, as in (ii). The above four methods are denoted as Resampling-SVM-RFE1, Resampling-SVM-RFE2, Resampling-Univariate1, Resampling-Univariate2, respectively.

Three other methods were representatives of stochastic variable selection algorithms. Three instantiations of KIAMB algorithm (Peña et al., 2007) were used. KIAMB was applied with Fisher's Z-test for continuous data (gene expression data) and G^2 test for discrete data (artificial simulated data), parameter $K = 0.8$, and three statistical thresholds $\alpha = 0.01$, $\alpha = 0.005$, and $\alpha = 0.001$ (denoted as KIAMB1, KIAMB2, KIAMB3, respectively). The first threshold was

[8] Delong's test (DeLong et al., 1988) was used to compare AUC point estimates in experiments with real and resimulated gene expression data where the response variable had two categories. McNemar's test (Everitt, 1977) was used to compare accuracies in experiments with simulated data where the response variable had more than two categories and AUC measure was not applicable.

used by inventors of the method in the paper that introduced it (Peña et al., 2007), while the latter two often lead to more parsimonious signatures without loss of predictivity based on prior experiments. A standard statistical threshold $\alpha = 0.05$ in most cases did not lead to termination of the algorithm, that is why it was not used in this work. To make experiments computationally tractable and robust to outlier runs of KIAMB, a 10 minute time limit was imposed for a single run of the algorithm (i.e., to extract one signature).

Finally, an Iterative Removal method (Natsoulis et al., 2005) was also applied. The implementation of this method used a signature extraction algorithm HITON-PC (Aliferis et al., 2008a; Aliferis et al., 2003) since it typically yields more compact signatures with predictivity comparable or better to the other gene selection methods (Aliferis et al., 2006a). Statistical comparison tests to compare predictivity of the signatures (DeLong et al., 1988; Everitt, 1977) were also utilized.

APPENDIX H

GENERATION OF RESIMULATED MICROARRAY GENE EXPRESSION DATA

The ability to produce realistic simulated data is a critical component of evaluating multiple signature identification algorithms in a systematic manner. In order to obtain large, realistic networks and data capturing the characteristics of human gene expression data, I applied a High-Fidelity Data Resimulation technique that generates synthetic data from a causal process that is induced from the real data and guarantees that the synthetic data is indistinguishable from the real data. The method and its application are briefly outlined below, more details can be found in (Aliferis and Statnikov, 2007).

The High-Fidelity Data Resimulation technique involves 6 steps[9]. *First*, a gene network is reverse-engineered from a real gene expression dataset. This step is performed by (a) obtaining an undirected graph by running HITON-PC algorithm for each gene and a phenotypic response variable, (b) orienting the graph using greedy search-and-score learning with Bach's metric (Bach and Jordan, 2003), and (c) learning densities of each gene and phenotypic response variable using SVM regression (Schölkopf et al., 1999) and classification (Vapnik, 1998), respectively. *Second*, synthetic data is generated from the above network using logic sampling (Russell and Norvig, 2003). *Third*, a power-law relationship between genes and their connectivity is examined in the simulated network (Barabasi and Bonabeau, 2003; Jeong et al., 2000). *Fourth,* a powerful classifier is applied to distinguish real from simulated data. The harder it is to perform this classification task, the better is the quality of resimulation. *Fifth*, Fisher's Z-test is used to ensure that statistical dependencies and independencies true in the real data are preserved in simulated data and vice-versa. *Sixth*, the existence of multiple maximally predictive and non-redundant signatures in simulated data is demonstrated empirically.

[9] Notice that steps 3-6 are used only for quality assurance purposes.

The above process was applied to 1,000 variables (999 randomly selected genes and a phenotypic response variable) from the 12,600 gene probes in the Affymetrix U95A array lung cancer gene expression data of (Bhattacharjee et al., 2001). The phenotypic response variable denotes whether a subject has adenocarcinoma or squamous cell carcinoma. Once the network was reverse-engineered (step 1), a set of 30,000 samples was generated from this network (step 2). The synthetic network and data passed validation steps 3-6. More details are given in (Aliferis and Statnikov, 2007).

APPENDIX I

CRITERIA FOR MICROARRAY GENE EXPRESSION DATASET ADMISSIBILITY AND PROTOCOL FOR QUALITY ASSURANCE AND PROCESSING

Recall that discovery and validation microarray gene expression datasets either originate from different laboratories or from different assay platforms. The following criteria for dataset admissibility are imposed in the independent-dataset experiments of this thesis: same phenotype and same or very similar patient population in both datasets, both datasets produced by microarray gene expression platforms from Affymetrix, sample size in discovery dataset ≥ 100, and sample size in discovery dataset $\geq$ sample size in validation dataset. Once candidate pairs of discovery and validation datasets that satisfy the above criteria are identified, I use the following quality assurance and processing procedure: (i) remove all patients/samples that are common between discovery and validation datasets (if applicable); (ii) for clinical outcome prediction tasks, remove censored patients/samples; (iii) if different microarray platforms are used, include only matching probes (obtained by using Affymetrix Array Comparison Spreadsheets: http://www.affymetrix.com/support/technical/comparison_spreadsheets.affx); (iv) ensure same or comparable normalization of both datasets; (v) verify presence of at least moderate predictive signal of the phenotype (>0.6 area under ROC curve) by using a signature based on all genes, and finally (vi) ensure same or statistically indistinguishable performance of the signature based on all genes when trained and tested by holdout validation in the discovery dataset and when trained in the discovery dataset and tested in the validation dataset. The last step is used to ensure that the populations of patients/samples are comparable between two datasets. To perform statistical testing in this step, a 95% confidence interval is built around each of the two point estimates[10] of

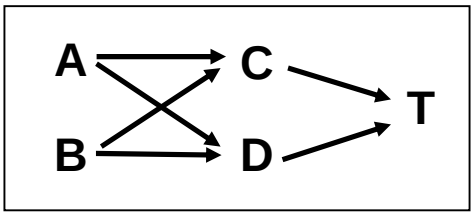
[10] One point estimate is obtained when a classifier is trained and tested by holdout validation in the discovery dataset, and another one is obtained when a classifier is trained in the discovery dataset and tested in the validation dataset.

area under ROC curve (DeLong et al., 1988) and it is verified that at least one of these confidence intervals includes a point estimate from another dataset.

APPENDIX J

AN EXAMPLE OF SIGNATURE MULTIPLICITY DUE TO SMALL SAMPLES

Consider a Bayesian network shown in Figure 20. It involves 5 variables including a response variable *T*. This network econdes a faithful distribution and thus only one Markov boundary exists in large samples, which is $\{C, D\}$. Now consider that one has access to three small samples from this distribution such that: in sample #1 one cannot reliably establish that $T \perp A \mid \{C, D\}$, in sample #2 one cannot reliably establish that $T \perp B \mid \{C, D\}$, and in sample #3 one cannot reliably establish either $T \perp A \mid \{C, D\}$ or $T \perp B \mid \{C, D\}$. Three Markov boundaries can be identified in the above samples, $\{C, D, A\}$, $\{C, D, B\}$, and $\{C, D, A, B\}$, respectively, assuming that neither *A* nor *B* significantly decreases the predictivity of *T* in given samples.



P($T \mid C, D$)	($C = 0, D = 0$)	($C = 0, D = 1$)	($C = 1, D = 0$)	($C = 1, D = 1$)
$T = 0$	0.2	0.5	0.7	0.4
$T = 1$	0.8	0.5	0.3	0.6

P($C \mid A, B$)	($A = 0, B = 0$)	($A = 0, B = 1$)	($A = 1, B = 0$)	($A = 1, B = 1$)
$C = 0$	0.3	0.7	0.9	0.4
$C = 1$	0.7	0.3	0.1	0.6

P($D \mid A, B$)	($A = 0, B = 0$)	($A = 0, B = 1$)	($A = 1, B = 0$)	($A = 1, B = 1$)
$D = 0$	0.6	0.7	0.8	0.4
$D = 1$	0.4	0.3	0.2	0.6

P(A)	
$A = 0$	0.6
$A = 1$	0.4

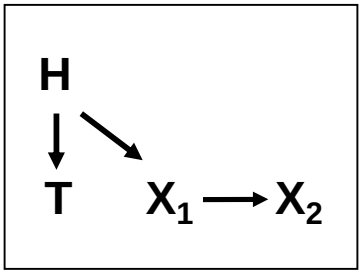
P(B)	
$B = 0$	0.4
$B = 1$	0.6

Figure 20: Graph of a Bayesian network used to illustrate signature multiplicity due to small sample sizes. The network parameterization is provided below the graph. The response variable is T. All variables take values {0, 1}.

APPENDIX K

AN EXAMPLE OF SIGNATURE MULTIPLICITY DUE TO HIDDEN VARIABLES

Consider a Bayesian network shown in Figure 21. It involves 4 variables including a response variable T. In the distribution with all variables observed, there is only one Markov boundary of T, which is $\{H\}$. Now consider that variable H is not observed. Because H is not observed and variables X_1 and X_2 contain exactly the same information about T, two Markov boundaries, $\{X_1\}$ and $\{X_2\}$, can be indentified in this distribution. Notice that all these Markov boundaries have reproducible but suboptimal (relative to the original distribution with H observed) predictivity of the response variable T.



P(T | H)	$H = 0$	$H = 1$
$T = 0$	0.9	0.2
$T = 1$	0.1	0.8

P(X_1 | H)	$H = 0$	$H = 1$
$X_1 = 0$	0.9	0.1
$X_1 = 1$	0.1	0.9

P(X_2 | X_1)	$X_1 = 0$	$X_1 = 1$
$X_2 = 0$	1.0	0.0
$X_2 = 1$	0.0	1.0

P(H)	
$H = 0$	0.3
$H = 1$	0.7

Figure 21: Graph of a Bayesian network used to illustrate signature multiplicity due to hidden variables. The network parameterization is provided below the graph. The response variable is T. All variables take values $\{0, 1\}$.

REFERENCES

Aliferis,C.F., Statnikov,A., Tsamardinos,I., Mani,S. and Koutsoukos,X.D. (2008a) Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part I: Algorithms and Empirical Evaluation. Accepted to the *Journal of Machine Learning Research*.

Aliferis,C.F. and Statnikov,A. (2007) High-Fidelity Resimulation from High-Throughput Data. *Technical Report DSL 07-03*.

Aliferis,C.F., Statnikov,A., Kokkotou,E., Massion,P.P. and Tsamardinos,I. (2006a) Local regulatory-network inducing algorithms for biomarker discovery from mass-throughput datasets. *Technical Report DSL 06-05*.

Aliferis,C.F., Statnikov,A., Pratap,S. and Kokkotou,E. (2006b) Statistical gene instability in gene-phenotype microarray association studies does not prohibit reproducibility: experimental evidence and network-theoretical justifications. *Technical Report DSL 06-06*.

Aliferis,C.F., Statnikov,A., Tsamardinos,I., Mani,S. and Koutsoukos,X.D. (2008b) Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part II: Analysis and Extensions. Accepted to the *Journal of Machine Learning Research*.

Aliferis,C.F., Tsamardinos,I. and Statnikov,A. (2003) HITON: a novel Markov blanket algorithm for optimal variable selection. *AMIA 2003 Annual Symposium Proceedings*, 21-25.

Azuaje,F. and Dopazo,J. (2005) *Data analysis and visualization in genomics and proteomics*. John Wiley, Hoboken, NJ.

Bach,F.R. and Jordan,M.I. (2003) Learning graphical models with Mercer kernels. *Advances in Neural Information Processing Systems (NIPS)*, **15**, 1009-1016.

Balazsi,G. and Oltvai,Z.N. (2007) A pitfall in series of microarrays: the position of probes affects the cross-correlation of gene expression profiles. *Methods Mol. Biol*, **377**, 153-162.

Barabasi,A.L. and Bonabeau,E. (2003) Scale-free networks. *Sci. Am.*, **288**, 60-69.

Beer,D.G., Kardia,S.L., Huang,C.C., Giordano,T.J., Levin,A.M., Misek,D.E., Lin,L., Chen,G., Gharib,T.G., Thomas,D.G., Lizyness,M.L., Kuick,R., Hayasaka,S., Taylor,J.M., Iannettoni,M.D., Orringer,M.B. and Hanash,S. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816-824.

Belsley,D.A., Kuh,E. and Welsch,R.E. (1980) *Regression diagnostics: identifying influential data and sources of collinearity*. Wiley, New York.

Bhattacharjee,A., Richards,W.G., Staunton,J., Li,C., Monti,S., Vasa,P., Ladd,C., Beheshti,J., Bueno,R., Gillette,M., Loda,M., Weber,G., Mark,E.J., Lander,E.S., Wong,W., Johnson,B.E., Golub,T.R., Sugarbaker,D.J. and Meyerson,M. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. U. S. A*, **98**, 13790-13795.

Bondell,H.D. and Reich,B.J. (2008) Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, **64**, 115-123.

Chu,T., Glymour,C., Scheines,R. and Spirtes,P. (2003) A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics*, **19**, 1147-1152.

Dave,S.S., Fu,K., Wright,G.W., Lam,L.T., Kluin,P., Boerma,E.J., Greiner,T.C., Weisenburger,D.D., Rosenwald,A., Ott,G., Muller-Hermelink,H.K., Gascoyne,R.D., Delabie,J., Rimsza,L.M., Braziel,R.M., Grogan,T.M., Campo,E., Jaffe,E.S., Dave,B.J., Sanger,W., Bast,M., Vose,J.M., Armitage,J.O., Connors,J.M., Smeland,E.B., Kvaloy,S., Holte,H., Fisher,R.I., Miller,T.P., Montserrat,E., Wilson,W.H., Bahl,M., Zhao,H., Yang,L., Powell,J., Simon,R., Chan,W.C. and Staudt,L.M. (2006) Molecular diagnosis of Burkitt's lymphoma. *N. Engl. J Med*, **354**, 2431-2442.

DeLong,E.R., DeLong,D.M. and Clarke-Pearson,D.L. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**, 837-845.

Dettling,M. and Buhlmann,P. (2002) Supervised clustering of genes. *Genome Biol*, **3**.

Dettling,M. and Buhlmann,P. (2004) Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis*, **90**, 106-131.

Dougherty,E. and Brun,M. (2006) On the number of close-to-optimal feature sets. *Cancer Informatics*, **2**, 189-196.

Dyrskjot,L., Zieger,K., Real,F.X., Malats,N., Carrato,A., Hurst,C., Kotwal,S., Knowles,M., Malmstrom,P.U., de la,T.M., Wester,K., Allory,Y., Vordos,D., Caillault,A., Radvanyi,F., Hein,A.M., Jensen,J.L., Jensen,K.M., Marcussen,N. and Orntoft,T.F. (2007) Gene expression signatures predict outcome in non-muscle-invasive bladder carcinoma: a multicenter validation study. *Clin. Cancer Res.*, **13**, 3545-3551.

Ein-Dor,L., Kela,I., Getz,G., Givol,D. and Domany,E. (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171-178.

Ein-Dor,L., Zuk,O. and Domany,E. (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. U. S. A*, **103**, 5923-5928.

Everitt,B. (1977) *The analysis of contingency tables*. Chapman and Hall, London.

Fawcett,T. (2003) ROC Graphs: Notes and Practical Considerations for Researchers. *Technical Report, HPL-2003-4, HP Laboratories*.

Freije,W.A., Castro-Vargas,F.E., Fang,Z., Horvath,S., Cloughesy,T., Liau,L.M., Mischel,P.S. and Nelson,S.F. (2004) Gene expression profiling of gliomas strongly predicts survival. *Cancer Res.*, **64**, 6503-6510.

Furey,T.S., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906-914.

Gold,D.L., Wang,J. and Coombes,K.R. (2005) Inter-gene correlation on oligonucleotide arrays: how much does normalization matter? *Am. J Pharmacogenomics.*, **5**, 271-279.

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

Grate,L.R. (2005) Many accurate small-discriminatory feature subsets exist in microarray transcript data: biomarker discovery. *BMC Bioinformatics*, **6**, 97.

Guyon,I., Aliferis,C.F. and Elisseeff,A. (2007) Causal Feature Selection. In Liu,H. and Motoda,H. (eds), *Computational Methods of Feature Selection*. Chapman and Hall.

Guyon,I., Gunn,S., Nikravesh,M. and Zadeh,L.A. (2006) *Feature extraction: foundations and applications*. Springer-Verlag, Berlin.

Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**, 1157-1182.

Guyon,I., Weston,J., Barnhill,S. and Vapnik,V. (2002) Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389-422.

Hadi,A.S. and Velleman,P.F. (1987) Comment: Diagnosing near collinearities in least squares regression. *Statistical Science*, **2**, 93-98.

Hammer,B. and Gersmann,K. (2003) A Note on the Universal Approximation Capability of Support Vector Machines. *Neural Processing Letters*, **17**, 43-53.

Hastie,T., Tibshirani,R., Botstein,D. and Brown,P. (2001) Supervised harvesting of expression trees. *Genome Biol*, **2**.

Hastie,T., Tibshirani,R., Eisen,M.B., Alizadeh,A., Levy,R., Staudt,L., Chan,W.C., Botstein,D. and Brown,P. (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol*, **1**.

Hollander,M. and Wolfe,D. (1999) *Nonparametric statistical methods*. Wiley, New York, NY, USA.

Ioannidis,J.P. (2005) Microarrays and molecular research: noise discovery? *Lancet*, **365**, 454-455.

Jeong,H., Tombor,B., Albert,R., Oltvai,Z.N. and Barabasi,A.L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651-654.

Jornsten,R. and Yu,B. (2003) Simultaneous gene clustering and subset selection for sample classification via MDL. *Bioinformatics*, **19**, 1100-1109.

Kohavi,R. and John,G.H. (1997) Wrappers for feature subset selection. *Artificial Intelligence*, **97**, 273-324.

Li,L., Weinberg,C.R., Darden,T.A. and Pedersen,L.G. (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, **17**, 1131-1142.

Meinshausen,N. (2008) Hierarchical testing of variable importance. *Biometrika*, **95**, 265-278.

Michiels,S., Koscielny,S. and Hill,C. (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**, 488-492.

Miller,L.D., Smeds,J., George,J., Vega,V.B., Vergara,L., Ploner,A., Pawitan,Y., Hall,P., Klaar,S., Liu,E.T. and Bergh,J. (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl. Acad. Sci. U. S. A*, **102**, 13550-13555.

Natsoulis,G., El,G.L., Lanckriet,G.R., Tolley,A.M., Leroy,F., Dunlea,S., Eynon,B.P., Pearson,C.I., Tugendreich,S. and Jarnagin,K. (2005) Classification of a large microarray data set: algorithm comparison and analysis of drug signatures. *Genome Res.*, **15**, 724-736.

Neapolitan,R.E. (2004) *Learning Bayesian networks*. Pearson Prentice Hall, Upper Saddle River, NJ.

Park,M.Y., Hastie,T. and Tibshirani,R. (2007) Averaged gene expressions for regression. *Biostatistics*, **8**, 212-227.

Pearl,J. (1988) *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers, San Mateo, California.

Pearl,J. (2000) *Causality: models, reasoning, and inference*. Cambridge University Press, Cambridge, U.K.

Peña,J., Nilsson,R., Björkegren,J. and Tegnér,J. (2007) Towards scalable and data efficient learning of Markov boundaries. *International Journal of Approximate Reasoning*, **45**, 211-232.

Phillips,H.S., Kharbanda,S., Chen,R., Forrest,W.F., Soriano,R.H., Wu,T.D., Misra,A., Nigro,J.M., Colman,H., Soroceanu,L., Williams,P.M., Modrusan,Z., Feuerstein,B.G. and Aldape,K. (2006) Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell*, **9**, 157-173.

Ploner,A., Miller,L.D., Hall,P., Bergh,J. and Pawitan,Y. (2005) Correlation test to assess low-level processing of high-density oligonucleotide microarray data. *BMC Bioinformatics*, **6**, 80.

Qiu,X., Brooks,A.I., Klebanov,L. and Yakovlev,N. (2005) The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics*, **6**, 120.

Ramaswamy,S., Ross,K.N., Lander,E.S. and Golub,T.R. (2003) A molecular signature of metastasis in primary solid tumors. *Nat Genet*, **33**, 49-54.

Ramsey,J., Zhang,J. and Spirtes,P. (2006) Adjacency-Faithfulness and Conservative Causal Inference. *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*.

Richardson,T. and Spirtes,P. (1999) Automated discovery of linear feedback models. In Glymour,C. and Cooper,G.F. (eds), *Computation, causation, and discovery*. MIT Press, Menlo Park, CA.

Richardson,T.S. and Spirtes,P. (2002) Ancestral graph Markov models. *Annals of Statistics*, **30**, 962-1030.

Roepman,P., Kemmeren,P., Wessels,L.F., Slootweg,P.J. and Holstege,F.C. (2006) Multiple robust signatures for detecting lymph node metastasis in head and neck cancer. *Cancer Res.*, **66**, 2361-2366.

Ross,M.E., Zhou,X., Song,G., Shurtleff,S.A., Girtman,K., Williams,W.K., Liu,H.C., Mahfouz,R., Raimondi,S.C., Lenny,N., Patel,A. and Downing,J.R. (2003) Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, **102**, 2951-2959.

Russell,S.J. and Norvig,P. (2003) *Artificial intelligence: a modern approach*. Prentice Hall/Pearson Education, Upper Saddle River, N.J.

Savage,K.J., Monti,S., Kutok,J.L., Cattoretti,G., Neuberg,D., De,L.L., Kurtin,P., Dal,C.P., Ladd,C., Feuerhake,F., Aguiar,R.C., Li,S., Salles,G., Berger,F., Jing,W., Pinkus,G.S., Habermann,T., la-Favera,R., Harris,N.L., Aster,J.C., Golub,T.R. and Shipp,M.A. (2003) The molecular signature of mediastinal large B-cell lymphoma differs from that of other diffuse large B-cell lymphomas and shares features with classical Hodgkin lymphoma. *Blood*, **102**, 3871-3879.

Schölkopf,B., Burges,C.J.C. and Smola,A.J. (1999) *Advances in kernel methods: support vector learning*. MIT Press, Cambridge, Mass.

Shawe-Taylor,J. and Cristianini,N. (2004) *Kernel methods for pattern analysis*. Cambridge University Press, Cambridge, UK.

Somorjai,R.L., Dolenko,B. and Baumgartner,R. (2003) Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, **19**, 1484-1491.

Sotiriou,C., Wirapati,P., Loi,S., Harris,A., Fox,S., Smeds,J., Nordgren,H., Farmer,P., Praz,V., Haibe-Kains,B., Desmedt,C., Larsimont,D., Cardoso,F., Peterse,H., Nuyten,D., Buyse,M., van,d., V, Bergh,J., Piccart,M. and Delorenzi,M. (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl. Cancer Inst.*, **98**, 262-272.

Spirtes,P., Glymour,C.N. and Scheines,R. (2000) *Causation, prediction, and search*. MIT Press, Cambridge, Mass.

Statnikov,A., Wang,L. and Aliferis,C.F. (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, **9**, 319.

Statnikov,A., Aliferis,C.F., Tsamardinos,I., Hardin,D. and Levy,S. (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, **21**, 631-643.

Stewart,G.W. (1987) Collinearity and least squares regression. *Statistical Science*, **2**, 68-84.

Su,A.I., Welsh,J.B., Sapinoso,L.M., Kern,S.G., Dimitrov,P., Lapp,H., Schultz,P.G., Powell,S.M., Moskaluk,C.A., Frierson,H.F., Jr. and Hampton,G.M. (2001) Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res.*, **61**, 7388-7393.

Tsamardinos,I. and Aliferis,C.F. (2003) Towards principled feature selection: relevancy, filters and wrappers. *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics (AI & Stats)*.

Tsamardinos,I., Aliferis,C.F. and Statnikov,A. (2003a) Algorithms for large scale Markov blanket discovery. *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, 376-381.

Tsamardinos,I., Aliferis,C.F. and Statnikov,A. (2003b) 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)*, 673-678.

Tsamardinos,I. and Brown,L.E. (2008) Markov Blanket-Based Variable Selection in Feature Space. *Technical report DSL-08-01*.

van de Vijver,M.J., He,Y.D., van't Veer,L.J., Dai,H., Hart,A.A., Voskuil,D.W., Schreiber,G.J., Peterse,J.L., Roberts,C., Marton,M.J., Parrish,M., Atsma,D., Witteveen,A., Glas,A., Delahaye,L., van,d., V, Bartelink,H., Rodenhuis,S., Rutgers,E.T., Friend,S.H. and Bernards,R. (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J Med*, **347**, 1999-2009.

Vapnik,V.N. (1998) *Statistical learning theory*. Wiley, New York.

Wagner,F. and Radelof,U. (2007) Performance of different small sample RNA amplification techniques for hybridization on Affymetrix GeneChips. *J Biotechnol*, **129**, 628-634.

Wang,L., Zhu,J. and Zou,H. (2006) The Doubly Regularized Support Vector Machine. *Statistica Sinica*, **16**, 589-615.

Wang,Y., Klijn,J.G., Zhang,Y., Sieuwerts,A.M., Look,M.P., Yang,F., Talantov,D., Timmermans,M., Meijer-van Gelder,M.E., Yu,J., Jatkoe,T., Berns,E.M., Atkins,D. and Foekens,J.A. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671-679.

Weissfeld,L.A. and Sereika,S.M. (1991) A multicollinearity diagnostic for generalized linear models. *Communications in Statistics-Theory and Methods*, **20**, 1183-1198.

Yeoh,E.J., Ross,M.E., Shurtleff,S.A., Williams,W.K., Patel,D., Mahfouz,R., Behm,F.G., Raimondi,S.C., Relling,M.V., Patel,A., Cheng,C., Campana,D., Wilkins,D., Zhou,X., Li,J., Liu,H., Pui,C.H., Evans,W.E., Naeve,C., Wong,L. and Downing,J.R. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133-143.

Zhang,J. and Spirtes,P. (2005) A Transformational Characterization of Markov Equivalence for Directed Acyclic Graphs with Latent Variables. *Proceedings of Uncertainty in Artificial Intelligence (UAI)*.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B(Statistical Methodology)*, **67**, 301-320.