BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 15 2010, pages 1887–1894
doi:10.1093/bioinformatics/btq261

Data and text mining

Advance Access publication June 2, 2010

Structure-based variable selection for survival data

Vincenzo Lagani[1,*] and Ioannis Tsamardinos[2]

[1]Institute of Computer Science, Foundation for Research and Technology - Hellas (FORTH) and [2]Computer Science Department, University of Crete, Heraklion, Greece

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Variable selection is a typical approach used for molecular-signature and biomarker discovery; however, its application to survival data is often complicated by censored samples. We propose a new algorithm for variable selection suitable for the analysis of high-dimensional, right-censored data called Survival Max–Min Parents and Children (SMMPC). The algorithm is conceptually simple, scalable, based on the theory of Bayesian networks (BNs) and the Markov blanket and extends the corresponding algorithm (MMPC) for classification tasks. The selected variables have a structural interpretation: if T is the survival time (in general the time-to-event), SMMPC returns the variables adjacent to T in the BN representing the data distribution. The selected variables also have a causal interpretation that we discuss.

Results: We conduct an extensive empirical analysis of prototypical and state-of-the-art variable selection algorithms for survival data that are applicable to high-dimensional biological data. SMMPC selects on average the smallest variable subsets (less than a dozen per dataset), while statistically significantly outperforming all of the methods in the study returning a manageable number of genes that could be inspected by a human expert.

Availability: Matlab and R code are freely available from http://www.mensxmachina.org

Contact: vlagani@ics.forth.gr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 10, 2010; revised on April 29, 2010; accepted on May 15, 2010

1 INTRODUCTION

Survival-analysis studies the occurrence and timing of events of interest. Often in medicine and biology, the event of interest is death, hence the name survival analysis; however, the field is much broader as the event of interest could be disease relapse or any other event. A typical characteristic of survival data is that they are often right-censored, i.e. the data may record subject i as alive (not having experienced the event) until time f_i (follow-up time) but may not contain the exact time t_i the event actually occurred. Excluding the censored data from the analysis systematically excludes the instances with higher probability of survival, and thus severely skews the results. Thus, typical variable selection and regression methods cannot be applied to model the time-to-event T. Unfortunately, it is often the case that analysis techniques cannot be trivially extended for survival data. Arguably, this is one reason why only a handful of variable selection methods are available for this type of data.

Variable selection's primary aim is often to improve *understanding* of the data-generating process. For example, a biologist may be more interested in the genes predictive of survival, than in the actual predictive model. Presumably, the variables involved in an optimally predictive, minimal, irreducible model carry unique information that provides insight into the data-generating process (we discuss this issue in more detail in the sections to follow). Given the above observations, algorithms with clear semantics for the selected variables, and optimality properties are desirable.

In this article, we adapt and evaluate an existing, successful and scalable method for variable selection called the Max–Min Parents and Children algorithm (MMPC; Tsamardinos *et al.*, 2003, 2006) to survival analysis; the method is based on the theory of Bayesian networks (BN) and Markov blanket-based variable selection (Tsamardinos and Aliferis, 2003) that has been recently shown to perform exceptionally well with complete (non-censored) data (Aliferis *et al.*, 2003, 2010). The Max–Min part of the name refers to the heuristic used in the algorithm regarding the order of consideration of the variables; the Parents and Children part of the name refers the structural interpretation of the selected variables: under certain conditions, they are the parents and children (adjacent nodes) of T in the BN representing the data distribution. We call the novel algorithm Survival MMPC (SMMPC). Under certain conditions, the selected nodes consist of an optimal set for prediction and minimum in terms of size. In other words, the selected variables contain no irrelevant (noise) or superfluous (redundant) variables, facilitating the interpretation of the model. For microarray data, the selected gene expressions always provide predictive information for survival, even when any other combination of gene expressions is known. Under certain conditions, this implies that no other gene is expected to be causally mediating the dependence. The set of parents and children is also connected to the set of direct causes of T; the causal interpretation of the returned variables is also discussed.

SMMPC is evaluated against a variety of prototypical and state-of-the-art variable selection algorithms that can scale up to the dimensionality often encountered in bioinformatics: Univariate Association Filtering, Forward Stepwise Selection, Bayesian Variable Selection (BVS) based on greedy search, Lasso Cox Regression (used as selection method) and Bayesian Selection based on Markov-Chain Monte-Carlo (MCMC). The variable selection methods are coupled with several survival regression methods. The evaluation is performed on several real, publicly available, high-dimensional, gene expression survival datasets. The results include

*To whom correspondence should be addressed.

Algorithm 1 SMMPC

1. **procedure** SMMPC(T, k, a)
   Input: T the target variable to predict, the maximum size of a conditioning set k, the threshold for rejecting independence a.
   Output: A subset of variables in $\mathcal{V}$
2. $\mathcal{R}:=\mathcal{V}$ % Remaining to consider
3. $\mathcal{S}:=\emptyset$ % Select so far
4. **repeat**
5. **if** $\exists X\in\mathcal{R}, \mathbf{Z}\subseteq\mathcal{S}, |\mathbf{Z}|\leq k$, s.t., $p_{XT|\mathbf{Z}}>a$ **then**
6. $\mathcal{R}:=\mathcal{R}\setminus\{X\}$
7. **end if**
   % Max–Min Heuristic
8. $M:=\arg\max_{X\in R}\min_{\mathbf{Z}\subseteq\mathcal{S}\setminus\{X\},|\mathbf{Z}|\leq k}(-p_{XT|\mathbf{Z}})$
9. $\mathcal{R}:=\mathcal{R}\setminus\{M\}$
10. $\mathcal{S}:=\mathcal{S}\cup\{M\}$
11. **until** $\mathcal{R}\neq\emptyset$
12.
13. $\forall X, s.t., \exists\mathbf{Z}\subseteq\mathcal{S}\setminus\{X\}, |\mathbf{Z}|\leq k$, $p_{XT|\mathbf{Z}}>c$
14. $\mathcal{S}:=\mathcal{S}\setminus\{X\}$
15. **return** $\mathcal{S}$
16. **end procedure**

an ordering of efficacy of both variable selection and regression methods as well as an analysis of the stability and learning curves of the methods.

2 SMMPC

We now present the SMMPC algorithm, and subsequently we discuss its theoretical properties and sufficient conditions for optimality and causal interpretation. The algorithm is essentially *MMPC without the symmetry correction*, as we call it, enhanced with a statistical test of conditional independence for censored data; it was first appeared in Tsamardinos *et al.* (2003) for complete data and in more detail in Tsamardinos *et al.* (2006). The algorithm accepts the target variable T and two tuning parameters, k the maximum size of conditioning set and a the level for accepting dependence.

We denote with $\mathcal{V}$ the predicting variables, T the time-to-event and C the time-to-censoring. What is observed is $F=\min(T,C)$, the follow-up time, i.e. the time for which we have been gathering information about the patient. The censoring status $\Delta=1(T<C)$ takes value 1 for data that are observed ($F=T$) and 0 for the cases are censored ($F=C$). Survival data $\mathcal{D}$ can be modeled as a set of m subjects indexed by i each represented by the triplet $\langle\mathbf{v}_i, f_i, \delta_i\rangle$, where $\mathbf{v}_i$ is the vector of n predicting variables, f_i the follow-up time and δ_i the censoring status.

The objective of SMMPC is to find the neighbors of T in a minimal BN representing the data distribution. Under certain conditions (see Section 3), this coincides with a minimum-size set of variables needed in order to optimally predict T. A basic component of SMMPC is a *test of conditional independence* using censored data. Let us denote with $\text{Ind}(X;T|\mathbf{Z})$, $X\in\mathcal{V}$, $\mathbf{Z}\subseteq\mathcal{V}$ the conditional independence of X with T given the variable set $\mathbf{Z}$ and with $\text{Dep}(X;T|\mathbf{Z})\equiv\neg\text{Ind}(X;T|\mathbf{Z})$ the corresponding dependence. The null hypothesis of the conditional independence test is that $\text{Ind}(X;T|\mathbf{Z})$ holds in the data distribution; the dependence is the alternative hypothesis. We denote with $p_{XT|\mathbf{Z}}$ the P-value provided by the test. If $p_{XT|\mathbf{Z}}\leq a$, where a is a user-defined threshold the null is rejected and we accept $\text{Dep}(X;T|\mathbf{Z})$. Otherwise, the null hypothesis of independence $\text{Ind}(X;T|\mathbf{Z})$ is accepted. The test of independence is the only place in the algorithm where the actual data are used.

Intuitively, it seems that *if a variable becomes independent of T given some $\mathbf{Z}$ it is superfluous for predicting T*, i.e. given $\mathbf{Z}$, X contains no additional information for T. SMMPC employs this basic premise to exclude from selection of all variables X for which it can find a $\mathbf{Z}\subseteq\mathcal{V}\setminus\{X\}$, s.t., $\text{Ind}(X;T|\mathbf{Z})$. The number of possible subsets for each variable is exponential and a brute-force approach is intractable. However, SMMPC employs BN theory to efficiently identify a $\mathbf{Z}$ for which $\text{Ind}(X;T|\mathbf{Z})$ if one exist under the conditions stated in Section 3, and make the algorithm practical for even very high-dimensional datasets. In practice, only the conditioning sets with $|\mathbf{Z}|\leq k$ are considered, because when $|\mathbf{Z}|$ is large, the statistical power of the tests is low. Prior experience with this type of algorithms (Aliferis *et al.*, 2010; Tsamardinos *et al.*, 2006) as well as the results in this article shows that most distributions are captured by sparse networks; this in turn implies that a small value of k is sufficient to filter out most variables.

The basic premise that if $\exists\mathbf{Z}$, s.t., $\text{Ind}(X;T|\mathbf{Z})$, then X is superfluous for predicting T *is not always correct*. Even if $\text{Ind}(X;T|\mathbf{Z})$, it could still be the case that $\text{Dep}(X;T|\mathbf{Z}\cup\{W\})$ for some other $W\in\mathcal{V}$ and X may become *necessary* for optimal prediction given additional variables. This case is discussed in detail in the sections to follow, where sufficient conditions for soundness are provided.

The pseudo-code of SMMPC is shown in Algorithm 1. It begins with a set of candidate variables to consider $\mathcal{R}$, initially set to all possible predicting variables $\mathcal{V}$. The algorithm also maintains a set of currently selected variables $\mathcal{S}$ initially set to the empty set. It then removes from $\mathcal{R}$ any variable that becomes independent of T condition on some subset $\mathbf{Z}\subseteq\mathcal{S}$. Subsequently, it heuristically selects and moves a member of $\mathcal{R}$ into $\mathcal{S}$ and continues until $\mathcal{R}=\emptyset$. As a final step SMMPC removes from $\mathcal{S}$ any variables X for which $\exists\mathbf{Z}\subseteq\mathcal{S}\setminus\{X\}$, s.t., $\text{Ind}(X;T|\mathbf{Z})$.

SMMPC selects as the next variable to include in the $\mathcal{S}$ the one with the maximum–minimum association with T over all subsets $\mathbf{Z}$ of $\mathcal{S}$ (hence the name Max–Min). Association of T with X given $\mathbf{Z}$ is measured by $-p_{XT|\mathbf{Z}}$, so the smaller the P-value the larger the association. Intuitively, the heuristic selects next *the variable that remains mostly associated with T despite our best efforts (i.e. after conditioning on all subsets of $\mathcal{S}$) to make the variable independent of T*. Recent theoretical results (Tsamardinos and Brown, 2008) also discovered another interesting interpretation: the Max–Min heuristic selects next *the variable that minimizes an upper bound on the false discovery rate of the output* (i.e. the false discovery rate of the identification of the Parents and Children set). The Max–Min heuristic is important both for the efficiency of the algorithm and the quality of the output but it does not affect the asymptotic correctness properties.

In terms of time complexity, assuming the size of the final output $\mathcal{PC}$ is about the maximum size of $\mathcal{S}$ in any iteration and some optimizations are in place (Tsamardinos *et al.*, 2006) the algorithm performs $O(|\mathcal{V}||\mathcal{PC}|^k)$ tests of independence. Thus, for problems with sparse structure (few neighbors of T), its complexity grows linearly to the number of variables.

2.1 Test of conditional independence

The null hypothesis $\text{Ind}(X;T|\mathbf{Z})$ of the conditional independence test implies that X is not necessary for predicting T when $\mathbf{Z}$ (and only $\mathbf{Z}$) is given; under this condition, the conditional independence test can be thought of as a procedure for selecting the best between two nested models: a model predicting T given $\mathbf{Z}\cup\{X\}$ and a model with only $\mathbf{Z}$ as covariates. We employ and evaluate the log-likelihood ratio and the local score test (Klein and Moeschberger, 2003), both on them based on the widely used Cox regression model (Cox, 1972). Extensive experimentations showed that the former test is able to produce better results than the local score test by selecting almost the same number of variables in average, thus we retained only the log-likelihood ratio test for successive experiments. Details about tests implementation and comparative experiments are described in the Supplementary Section 1.

3 THE THEORY AND INTERPRETATION OF SMMPC

For a practitioner that employs variable selection for gaining insight and understanding in a domain, it is important to know in depth the theoretical properties and the semantic interpretation of the variables selected by an

algorithm. SMMPC, while conceptually simple, is based on relatively well-understood BN and Causal BN theory that can be used to analyze its properties and semantics. An expanded and more in-depth version of this section is included in the Supplementary Section 2.

3.1 BN structural interpretation

MMPC and its extension, SMMPC, return variables with a specific structural interpretation in the BN representing the data distribution. We do not distinguish between the two versions of the algorithm when not necessary. We briefly review the related BN theory. A BN $\langle G,P\rangle$ is a directed acyclic graph $G=\langle \mathcal{V},E\rangle$ and a probability distribution P, both defined on the set of random variables (nodes) $\mathcal{V}$. In the remaining of this section, we assume $\mathcal{V}$ is the full set of variables, including T. G and P are such that the *Markov condition* is satisfied: every variable is independent of any subset of its non-descendant variables conditioned on its parents (Pearl, 2000). A faithful BN is one where in addition, only the independencies that are entailed by the Markov condition hold in P. In other words, the independencies stem directly from the structure (graph) of the network and are not accidental properties of the parameters. Distributions P for which there exist a faithful network $\langle G,P\rangle$ are also called faithful. When choosing parameters randomly for a given structure, the probability of obtaining a non-faithful distribution has Lebesgue measure zero and so one expects them to be 'rare' (although this may not hold in systems obtained by Natural Selection; Brown and Tsamardinos, 2008).

For a given distribution P, there may be many graphs G such that the BN $\langle G,P\rangle$ is faithful. However, the set of neighbors of a variable T (parents and children of T) is unique in all networks faithful to the same distribution and denoted by $\mathcal{N}(T)$. Let us also define the extended neighbor set of order k, denoted with $\mathcal{EN}_k(T)$ as the set of variables that are dependent on T conditioned on any subset of the $\mathcal{N}(T)$ of size at most k. It is easy to see that $\mathcal{EN}_{k+1}(T)\subseteq\mathcal{EN}_k(T)$. We will denote with $\mathcal{EN}(T)$ the unrestricted cases for infinite k. In faithful distributions, it holds that $\mathcal{N}(T)\subseteq\mathcal{EN}_k(T)$ for any k. In Aliferis *et al.* (2010), we show the following:

PROPOSITION 1. *Assume that: (a) The data distribution P is faithful, i.e. there exist a faithful BN $\langle G,P\rangle$; (b) The conditional tests of independence make no statistical errors at level a; then*

$$\mathcal{N}(T)\subseteq \text{MMPC}(T,k,a)\subseteq\mathcal{EN}_k(T).$$

It is possible that MMPC outputs variables not in $\mathcal{N}(T)$ under the assumptions, i.e. $\mathcal{N}(T)\subset MMPC(T,\infty)$. First, we note that this type of *false positive* has to be a descendant of T in the network; second, they can be removed to end up with the true $\mathcal{N}(T)$ by performing additional executions of MMPC and by employing what is called the 'symmetry correction'. In Aliferis *et al.* (2010), it was determined empirically that the overhead for the symmetry correction does not outweigh the theoretical benefits and that these cases are rare. Thus, we do not employ the symmetry correction in the algorithm and we will assume that $\mathcal{N}(T)=\mathcal{EN}(T)$ and so MMPC(T,∞,a) outputs $\mathcal{N}(T)$ under the conditions above.

3.2 Optimality properties and relation to the Markov blanket

In this section, we examine conditions under which the selected variable subset $\mathcal{N}(T)$ is optimally predictive and minimum-in-size. An emerging and principled approach in variable selection is based on identifying the Markov blanket of the variable T to predict (hereafter, the target variable). The Markov blanket of T (denoted as $\mathcal{MB}(T)$) relative to a set of measured variables $\mathcal{V}$ is defined as a minimal set conditioned on which all other variables in $\mathcal{V}$ become independent of T [this is the Markov Boundary in the terminology of Pearl (2000)]: $P(T|\mathcal{V}\setminus\{T\})=P(T|\mathcal{MB}(T))$. Thus, all information for optimally predicting T is contained within the $\mathcal{MB}(T)$ and, therefore, it may seem that this is a minimal set of variables required for optimal prediction. The latter statement is not true in general, however, as the learner and the performance metric used are important. For the $\mathcal{MB}(T)$ to be the solution to the variable selection problem as it was defined above, two conditions are sufficient (Tsamardinos and Aliferis, 2003):

(1) *The learner used to construct the prediction model can correctly estimate the distribution $P(T|\mathcal{MB}(T))$.*

(2) *The performance metric is such that perfect estimation of the conditional probability distribution of T is required with the smallest number of variables possible.*

The above conditions often hold or hold approximately in many typical applications. Specifically in our experiments, we use several regression methods and optimize their parameters trying to find the one that best matches the distribution in order to satisfy Condition (1). In addition, the performance metrics used [Concordance Index (CI) and Integrated Brier Score (IBS)] are optimized only when full knowledge of T's conditional distribution is optimized.

Identifying the $\mathcal{MB}(T)$ in the general case is a difficult problem. New theoretical results connect the $\mathcal{MB}(T)$ with the structure of the BN capturing the data distribution and gave rise to time and sample-efficient algorithms for identifying the Markov blanket in faithful distributions represented by sparse networks; such algorithms include the Max–Min Markov blanket (MMMB; Tsamardinos *et al.*, 2003) and the HITON (Aliferis *et al.*, 2003) algorithms.

We now briefly present some of these results. Recall that $\mathcal{N}(T)$ denotes the set of neighbors of T in any graph G faithful to the data distribution, while we use $\mathcal{S}(T)$ to denote the set of all parents of common children of T not in $\mathcal{N}(T)$. By definition, $\mathcal{N}(T)\cap\mathcal{S}(T)=\emptyset$. In a faithful distribution it can been shown that $\mathcal{MB}(T)$, $\mathcal{N}(T)$ and $\mathcal{S}(T)$ are unique. Thus, $\mathcal{MB}(T)$ is not only minimal but also minimum; in addition, the Markov blanket has a graphical interpretation and specifically

$$\mathcal{MB}(T)=\mathcal{N}(T)\cup\mathcal{S}(T)$$

i.e. *the Markov blanket of T is the set of parents, children and parents of common children with T*.

In the Section 3.1, we show that SMMPC returns $\mathcal{N}(T)$ under the assumptions stated, a subset of the Markov blanket of T. Let us now turn our attention to the remaining part, the $\mathcal{S}(T)$ variables. Recall, the basic premise for selecting variables using MMPC is that if a variable X becomes independent of T given some $\mathbf{Z}$ it is superfluous for predicting T. In fact, *variables in $\mathcal{S}(T)$ are exactly the variables for which this assumption fails*: for an $X\in\mathcal{S}(T)$, there exist a subset $\mathbf{Z}$ for which Ind$(X;T|\mathbf{Z})$ (otherwise X would have an edge to or from T according to the theorem) but X is in $\mathcal{MB}(T)$, and thus required for optimal prediction. There do exist techniques to identify $\mathcal{S}(T)$ but would require more complicated method for tests of conditional independence with the censored-variable T as a covariate; as a first attempt at structure-based variable selection with censored data, SMMPC performs variable selection by approximating the Markov blanket of T with $\mathcal{N}(T)$.

3.3 Causal interpretation

In recent work (Aliferis *et al.*, 2010), we have shown that in simulated studies with known causal structure, prominent-variable-selection algorithms indeed select subsets that lead to models with close-to-optimal predictive performance; unfortunately, however, the selected variables are distributed almost at random on the causal graph and so they could not be used to understand the domain. One could argue they are actually misleading. BN-based algorithms on the other hand, such as MMPC, output variables that also have a specific causal interpretation, under certain conditions discussed below.

We will first assume the standard causal discovery setting of Pearl (2000). That is, we assume that there exists a *Causal* BN defined on the observed variables $\mathcal{V}$, that faithfully captures the data distribution. In that network, an edge $X\rightarrow T$ means that X is *directly causing* (affecting) the survival time: if one manipulates X (e.g. increase medication), a change in the distribution of T will be observed. In this case, the $\mathcal{N}(T)$ is the set of neighbors of T in the causal network (and in any other network faithful to the distribution) and so

it is the set of direct causes and direct effects of T. Since in our setting of survival analysis all variables are measured prior to T, the output of MMPC in this case can be interpreted as the set of direct causes of survival. In addition, since T has no causal effects, $\mathcal{S}=\emptyset$ and so $\mathcal{MB}(T)=\mathcal{N}(T)$. In summary,

PROPOSITION 2. *Assume that T is the time-to-event, thus having no causal effects in $\mathcal{V}$, (a) the data distribution P is faithful to a* Causal BN*, i.e. there exist a faithful* and Causal *BN $\langle G,P\rangle$, (b) the conditional tests of independence make no statistical errors at level a then SMMPC(T,∞,a) outputs the* direct causes *of T (direct in the context of the remaining variables), which is also the Markov blanket of T.*

The assumption of the distribution being faithful to a Causal BN has at least three subtle and implicit subparts. First, the causal mechanism is assumed acyclic, thus there should be no feedback loops. Second, a causal interpretation of the Markov condition (known as the Causal Markov condition) may be violated due to measurement noise. Third, the assumption of a faithful Causal BN for the distribution implies no hidden confounders, i.e. latent variables that are causes of two modeled variables. This assumption is called *Causal Sufficiency* in the terminology of Spirtes *et al.* (2000). The implications of violations of these assumptions are discussed in detail in the Supplementary Material, as well as techniques to relax the assumptions. While possible violations of the assumptions require care in interpreting the output, we note that non-causally based algorithms fail to return causally meaningful variables even in simple simulated problems, where all of the above standard assumptions hold by design (Aliferis *et al.*, 2010).

4 REVIEW OF RELATED WORK

Survival data analysis has been an active field in statistics for decades and dozens of regression algorithms have appeared in the literature. The recent emergence of high-dimensional, biological datasets presents new challenges to all aspects of analysis (see van Wieringen *et al.*, 2009; Witten and Tibshirani, 2009 for a review of recent methods). A standard approach to deal with the high-dimensionality is to search for models that attempt to minimize both the error and the number of the coefficients of the variables, e.g. Survival Trees (Hothorn *et al.*, 2004), Ridge Cox Regression (Hoerl and Kennard, 2000) and Lasso Cox Regression (Tibshirani, 1997a). Another approach to address high-dimensional, survival analysis is to first reduce the dimensionality of the problem before applying regression. Dimensionality reduction methods for survival analysis tasks in bioinformatics include clustering algorithms (Hastie *et al.*, 2001), principal component analysis (Li and Gui, 2004) or partial least-square procedures (Nguyen and Rocke, 2002); see Nguyen and Rojo (2009) for an empirical comparison among different dimension reduction algorithms. General dimensionality reduction often transforms the data to lower dimensional spaces that are hard to interpret in terms of the original, input variables and measured quantities. Thus, dimensionality reduction is not appropriate in general when the goal of the analysis is to understand the effect of the measured quantities to the outcome variable. A restricted form of dimensionality reduction is variable selection, where the data are projected to a lower dimension space consisting of a subset of the original variables. Variable selection for regression and classification has been an intense field of study with hundreds of published algorithms, conferences and competitions (Guyon *et al.*, 2004). Unfortunately, it is often the case that these algorithms cannot trivially be adapted to censored, survival data; thus, the number of available variable selection algorithms for high-dimensional, survival data is relatively low.

5 COMPARATIVE EVALUATION

The comparative evaluation aims to address several questions of interest regarding regression and variable selection algorithms for high-dimensional, survival analysis tasks in bioinformatics. The primary goal of the evaluation is to validate the advantages of SMMPC over the prior state-of-the-art in variable selection for high-dimensional, survival analysis tasks in bioinformatics. Each variable selection algorithm is coupled with several regression algorithms to identify the best combination. Interesting conclusions are also derived about the performance of regression algorithms and their interaction with variable selection methods.

5.1 Variable selection algorithms

We consider general dimensionality reduction methods as out of the scope of this article, and focus on variable selection algorithms for high-dimensional survival data. We try to include as many methods as possible that have been applied to high-dimensional biological data. In addition, we found the BVS method in Faraggi and Simon (1998) interesting and intriguing; we have made simple adaptations to it so as to be applicable to high-dimensional data. A list of algorithms now follows:

5.1.1 No Selection We include No Selection (NS) (retain all original variables) as a baseline to compare against the performance of other methods. No parameter needs to be specified.

5.1.2 SMMPC This is the procedure introduced in Section 2. The parameters to set (see Algorithm 1) are k the maximum-size conditioning set allowed and a the significance level threshold for rejecting conditional independence. The parameter values have been optimized within $k\in\{2,3\}$ and $a\in\{0.05,0.10,0.15\}$.

5.1.3 Univariate Selection The variables are ranked in descending order of pairwise association with the time-to-event T, and the most associated variables are selected. We used the implementation by Bovelstad *et al.* (2007). The selected variables are the ones with $p_{XT|\emptyset}\leq a$, where a is a threshold provided as a parameter. The parameter value has been optimized within $a\in\{0.05,0.10,0.15\}$.

5.1.4 Forward Selection The method begins with the empty set as the selected predictors $\mathcal{S}$. In each subsequent step, it adds the variable X that maximizes the association of T conditioned on (i.e. in the context of) $\mathcal{S}$, i.e. minimizes $p_{XT|\mathcal{S}}$. We used the implementation by Bovelstad *et al.* (2007); this specific implementation of Forward Selection (FS) continues to add variables as long as $|\mathcal{S}|\leq\pi\cdot m$, where π is a parameter and m the number of samples. The parameter value has been optimized within $\pi\in\{0.01,0.05,0.1\}$ (the default value of the released code is 0.05).

5.1.5 BVS Prime The original BVS method first appeared in Faraggi and Simon (1998); unfortunately, this method requires performing intensive matrix operations, and does not scale 'as is' to high-dimensional data. Hence, in our implementation we introduced some minimal modifications in order to make the algorithm applicable to the tasks in our study: (a) perform Univariate Selection (US) to reduce the number of variables to consider; select only the top N variables (we use $N=100$ in our experiments), (b) a FS procedure substitutes the backward-selection procedure, (c) the ridge Cox regression (see Supplementary Section 3) was used instead of the Cox regression, due to its robustness to high-dimensional data; a fixed penalty term $\lambda=0.001$ is used to fit the model. We name the resulting method BVS Prime (BVS′). The parameter σ that we optimize is the hyperparameter corresponding to the SD of the priors of the coefficients. It is optimized within the set $\{0.1,0.5,1\}$ (the authors suggest values between 0.1 to 0.5).

5.1.6 MCMC selection This method is also based on Bayesian statistics and is introduced in Sha *et al.* (2006). It defines the prior distribution of the coefficients $\mathbf{b}$ as well as $\mathcal{S}$, the selected variables over an Accelerated Failure Time (AFT) model. It then employs the Metropolis–Hastings method (Hastings, 1970) to sample from the posterior distribution of $\mathbf{b}$ and $\mathcal{S}$. As the authors suggest, we select the most probable a posteriori $\mathcal{S}$ as the output of the method. We employ the implementation provided by the authors.

Table 1. Datasets used in the evaluation

Name and Reference	#Cases	#Cens	#Vars	Event
Vijver (van de Vijver *et al.*, 2002)	295	207	70	metastasis
Veer (van't Veer *et al.*, 2002)	78	44	4751	metastasis
Ros.2002 (Rosenwald *et al.*, 2002)	240	102	7399	survival
Ros.2003 (Rosenwald *et al.*, 2003)	92	28	8810	survival
Bullinger (Bullinger *et al.*, 2004)	116	49	6283	survival
Beer (Beer *et al.*, 2002)	86	62	7129	survival

The columns are in order, literature reference, number of training cases, number of censored cases, number of predicting variables and the event of interest.

MCMC requires the specification of a large number of parameters; we set all but one to their default values; we specify c the expected number of selected variables to the number selected by SMMPC on the same task (see Supplementary Section 7 for further details).

5.1.7 Lasso Selection First proposed in Tibshirani (1997b), the Lasso algorithm adds a penalty term to the log partial likelihood of the Cox regression model: $L(\mathbf{b}) - w||\mathbf{b}||_1$. Penalizing the $L1$-Norm shrinks the coefficients of redundant variables towards zero; thus, variables to be eliminated can be easily identified. The major drawback of this method is that it requires the solution of a non-derivable optimization problem, leading to elevated computational times. We used the implementation provided by Sohn *et al.* (2009), that is considerably faster than other freely available codes. The only parameter to be optimized is w, i.e. the weight that regulates the influence of the penalty term; we varied its values within the set $\{0.1, 1, 10\}$ (1 was the default value of the used implementation).

5.2 Regression algorithms

As demonstrated in Tsamardinos and Aliferis (2003) and Kohavi and John (1997), the interplay between variable selection method and the learner is important to identify the smallest variable subset with optimal predictive power. Thus, we couple each variable selection method with several regression methods in order to evaluate their performance. In particular, we employ Cox regression, Ridge Cox regression (Hoerl and Kennard, 2000), AFT models, Random Survival Forest (RSF; see Breiman and Schapire, 2001) and Support Vector Machine Censored Regression (SVCR; Shivaswamy *et al.*, 2007), optimizing the performances of each regressor on a variety of parameters. We exclude Lasso regression (as a regressor, not as a variable selection method) as dominated by Ridge Cox Regression in this domain and task (Bovelstad *et al.*, 2007). Supplementary Section 3 reports an exhaustive description of regressors and parameters that are optimized.

5.3 Dataset description

We identify several micro-array, gene expression, independent public datasets of survival studies with censored outcomes used in prior comparative studies (Bair and Tibshirani, 2004; van Wieringen *et al.*, 2009). Summary statistics and references are reported in Table 1. For all but the last dataset, we consider the data as preprocessed by the authors; we log-normalize the data of the last dataset (Beer *et al.*, 2002).

5.4 Performance metrics

Metrics of performance in the case of survival data is more complicated due to censorship: the error can only be computed exactly if the case is not censored. Thus, several specially designed evaluation metrics have been proposed in the literature, such as the Time-Depending area under the curve (AUC) (Heagerty *et al.*, 2000), the Weighted Classification Accuracy (Ripley and Ripley, 1998), the CI (Harrel, 2001) and the IBS (Graf *et al.*, 1999). We employ the latter two in this study.

The CI measures the percentage of pairs of subjects correctly ordered by the model in terms of their expected survival time. Notably, we can determine $t_i > t_j$ only when $f_i > f_j$ and $\delta_j = 1$, and similarly for the reverse relation. The pairs for which neither $t_i > t_j$ nor $t_i < t_j$ can be determined are excluded from the calculation of CI. *When there are no censored data, the CI is equivalent to the Area Under the Receiving Operating Characteristic curve*. Thus, a model ordering pairs at random (without use of the predicting variables) is expected to have a CI of 0.5, while perfect predictions would lead to a CI of 1. The Brier Score BS(t) measures the squared difference between the predicted survival probability and the observed outcome at time t, weighted for the loss of information due to the presence of censorship. The IBS consists in $\mathrm{IBS} = \max(T)^{-1} \int_0^{\max(T)} BS(t)dt$. IBS measures the estimation error assigning a value of 0 when the distribution of $S(T|v_i)$ is exactly estimated, while IBS = 0.25 is expected when predictions are random.

5.5 Parameter optimization and estimation of performance

A common procedure for parameter optimization is the use of a hold-out validation set. Each parameter combination is employed to learn a model on the training set and its performance is measured on the validation set. The best-performing parameter combination is then selected and a final model is learnt over all data (training plus validation). The best performance on the validation set is the maximum performance observed, and so it follows an extreme distribution. If a number of parameter combinations are attempted it is likely that the maximum observed performance is optimistic (the estimation is biased upwards; see Jensen and Cohen, 2000 for a discussion). Thus, a second hold-out set, a test set, is required for an unbiased estimation of performance. A drawback of the above protocol is that a portion of the data is not used for training. This is partially overcome by generalizing this procedure using cross-validation to what is known as nested N-fold cross-validation (Aliferis *et al.*, 2010; Dudoit and van der Laan, 2005; Statnikov *et al.*, 2005). An outer cross-validation loop considers several test sets for estimation of performance. For each one of them, an inner cross-validation loop considers several validation sets (and corresponding training sets) to fit models, select the best parameters and fit a model on the train-validation data using the best parameters. Notice that, the test data are only used for performance estimation and never to fit a model or select parameters. We also note that parameter optimization in the inner loop depends on the metric used each time, i.e. CI or IBS.

5.6 Determining statistical significance using permutation testing

We employ permutation testing as a non-parametric way to determine the statistical significance of the difference in performance between two methods. Let us define M_{1ij} and M_{2ij}, the models produced by the two methods, respectively, on dataset i when the test set is fold j. Let us call Π^0_{1ij} and Π^0_{2ij}, the set of predictions of the models on the test fold j of dataset i. We define our test statistic Σ_i to be the average difference of performance between Methods 1 and 2 over all folds j of dataset i. We define as the null hypothesis H_i the hypothesis that the $E(\Sigma_i) = 0$. Under the null hypothesis, it does not matter whether the predictions come from Π^0_{1ij} or Π^0_{2ij}. We thus create 1000 permuted sets of predictions Π^k_{1ij} and Π^k_{2ij}, $k = 1, \ldots, 1000$ produced by randomly swapping with probability 50% each pair of corresponding predictions $\pi_1 \in \Pi^0_{1ij}$ and $\pi_2 \in \Pi^0_{2ij}$. The test statistic Σ^k_i is calculated on the corresponding permuted prediction sets. The empirical distribution of $\{\Sigma^k_i\}$ estimates the distribution of Σ_i under the null hypothesis. The P-value of H_i is estimated as the percentage of times $|\Sigma^0_i| \leq |\Sigma^k_i|$, where Σ^0_i is the statistic calculated on the original (non-permuted) sets of predictions. We also define the null hypothesis that Method 1 produces more predictive models on dataset i than Method 2, denoted by $H_{1<2:i}$. Its P-value is simply the one-sided tail of the distribution, i.e. the percentage of times $\Sigma^0_i \leq \Sigma^k_i$. Similarly, we define the statistic Σ as the

Table 2. Nested cross-validated performances of feature selection methods

		NS		US		LS		FS		BVS′		SMMPC	
CI	Vijver	0.715	(0.017)	$^{-}$0.717	(0.016)	$^{-}$0.717	(0.036)	$^{---}$0.733	(0.055)	0.715	(0.009)	0.709	(0.012)
	Veer	$^{+}$0.676	(0.014)	0.690	(0.010)	0.686	(0.050)	$^{++}$0.648	(0.037)	$^{+++}$0.616	(0.017)	0.707	(0.075)
	Ros.2002	$^{-}$0.628	(0.032)	$^{-}$0.627	(0.025)	$^{++}$0.605	(0.038)	$^{+++}$0.586	(0.029)	$^{---}$0.638	(0.016)	0.618	(0.011)
	Ros.2003	0.724	(0.049)	0.707	(0.037)	$^{+++}$0.602	(0.414)	$^{++}$0.667	(0.049)	0.722	(0.064)	0.707	(0.058)
	Bullinger	$^{-}$0.639	(0.006)	0.633	(0.012)	$^{---}$0.669	(0.023)	$^{++}$0.593	(0.015)	$^{--}$0.647	(0.089)	0.617	(0.023)
	Beer	$^{-}$0.770	(0.095)	0.755	(0.067)	$^{+}$0.672	(0.045)	$^{+}$0.669	(0.133)	$^{++}$0.667	(0.114)	0.720	(0.118)
nVars	Vijver	70	(0.000)	44.3	(2.50)	9.75	(10.5)	20.65	(12.8)	10.5	(3.15)	6.15	(0.250)
	Veer	4751	(0.000)	1117	(254.3)	8.65	(5.55)	5.0	(3.15)	5.65	(0.25)	6.05	(0.5)
	Ros.2002	7399	(0.000)	1620	(445.5)	13.65	(2.75)	16.9	(8.5)	24.15	(4.875)	12.55	(1)
	Ros.2003	8810	(0.000)	1681	(211.6)	2.15	(1.375)	4.15	(1.125)	11.65	(1.65)	8.4	(0.75)
	Bullinger	6283	(0.000)	1230	(139.4)	6.15	(14.65)	6	(2.75)	16.4	(1.375)	8.05	(0.375)
	Beer	7129	(0.000)	1146	(99.2)	10.05	(4.625)	4.5	(1.5)	8.5	(0.125)	7.65	(2.15)
Mean	CI	$^{-}$0.692	(0.036)	0.688	(0.028)	$^{+++}$0.658	(0.101)	$^{+++}$0.649	(0.053)	$^{++}$0.668	(0.051)	0.680	(0.049)
	IBS	0.165	(0.022)	0.165	(0.010)	0.167	(0.045)	$^{+}$0.174	(0.019)	0.170	(0.041)	0.169	(0.031)
	nVars	5740	(0.00)	1140.0	(192.1)	8.4	(6.575)	9.53	(4.97)	12.8	(1.9)	8.14	(0.82)

The metrics are presented for the best-performing regression method; the numbers in parentheses show the range of the metric between the best and the worst regression method tried. Symbols $^{+}$, $^{++}$ and $^{+++}$: SMMPC outperforms the corresponding method at the significance levels of 0.1, 0.05 and 0.01, respectively; $^{-}$, $^{--}$ and $^{---}$: SMMPC is outperformed by the corresponding method at the significance levels of 0.1, 0.05 and 0.01, respectively.

average Σ_i over all datasets and the null hypothesis $H_{1<2}$ that Method 1 produces models of higher performance than Method 2 on average in all the data populations in the study. The P-values of this hypothesis are estimated in a similar way.

Regarding the prediction sets Π_{1ij} and Π_{2ij}, when the performance metric is the CI they consist of relative risk predictions $RR(p,q)$ for each unordered pair p,q of patients in the test set ($RR(p,q)=1$, if subject p is given a higher probability to survive longer than subject q). During permutations a relative risk for p,q as given by Method 1 may be swapped with a relative risk for p,q as given by Method 2. When the performance metric is the IBS, the prediction sets contain the patient-specific survival functions over all patients in the test set. During permutations a survival function for patient i as estimated by Method 1 may be swapped by the survival function for patient i given by Method 2.

6 RESULTS

Experimentation results are now provided. Since the results in terms of CI and IBS are substantially in agreement, we report only CI performances. IBS performances as well as adjunctive results are reported in the Supplementary Section 8.

6.1 Comparing variable selection methods

In this set of experiments, we compare the variable selection methods, namely NS, FS, Lasso Selection (LS), MCMC, BVS′ and SMMPC against each other. The MCMC procedure is so computationally costly that forced us to distinguish it from all other methods and use a simplified experimentation protocol to compare against. The details are in the Supplementary Section 7. SMMPC achieves a statistically significantly higher performance for both CI and IBS metrics, on most datasets individually as well as on all datasets collectively. Within the scope of our evaluation (particularly, the fact that we have used 50 000 iterations for the MCMC procedure), we consider the method dominated by SMMPC both in terms of computational efficiency and predictive performance of the resulting models.

To compare against all other methods, the nested cross-validation procedure is employed for parameter optimization and estimation of performance. The results are shown in Table 2 for the regression method that best matches the variable selection method. SMMPC serves as the baseline to compare against and produce the P-values of the permutation tests.

NS and US produce the highest performing models, even though the difference from SMMPC is not statistically significant for US and only marginally significant for NS ($P=0.052$) considering the CI metric. However, this performance is achieved at the expense of employing a quite large number of variables in the models and thus, these methods are not suitable for knowledge discovery tasks. To further clarify the relation between US and SMMPC, we repeated the experiments for US restricting the method to only select at most the top f variables, where f is the number of variables selected by SMMPC on that fold. This restricted version is outperformed by SMMPC at a statistical level of 0.01 and 0.05 for CI and IBS, respectively, showing the SMMPC orders the variables better than US in terms of the cumulative predictive information they carry.

We consider the remaining methods, namely FS, BVS′, LS, and SMMPC suitable for knowledge discovery, as they reduce the number of variables to a human-management size of less than a dozen. Comparing these methods on individual datasets does not provide a clear picture. However, over all datasets SMMPC produces higher performing models in terms of CI than all other methods ($P<0.05$). With respect to IBS, the same conclusions hold against FS and BVS′ even though the significance levels increase. In terms

Table 3. Statistical significance of comparing CI performance between each pair of variable selection methods

Methods	NS	US	SMMPC	BVS′	LS	FS
NS		0.21	0.052	<0.01	<0.01	<0.01
US			0.147	<0.01	<0.01	<0.01
SMMPC				0.049	<0.01	<0.01
BVS′					0.203	0.011
LS						0.104

Each cell reports the *P*-value for accepting the hypothesis the row-method outperforms the column method. Significant *P*-values at the level of 0.1 are underlined.

Table 4. Best regression method for each combination dataset and feature selection algorithm (CI metric)

Methods	NS	US	SMMPC	BVS′	LS	FS
Vijver	RSF	Ridge	RSF	SVCR	SVCR	SVCR
Veer	Ridge	SVCR	SVCR	Cox	RSF	Cox
Ros.2002	Ridge	Ridge	RSF	Cox	RSF	Cox
Ros.2003	Ridge	Ridge	Ridge	AFT	RSF	Ridge
Bullinger	Ridge	RSF	Ridge	SVCR	Ridge	RSF
Beer	Ridge	Ridge	SVCR	Ridge	RSF	AFT

of the size of the selected variable set, SMMPC selects on average the fewest number of variable, followed closely by LS. The range of performance between the best and the worst model (regressor) produced by each method is higher for the methods that select small variable sets (FS, LS, BVS′ and SMMPC). *This implies that for such methods it becomes more important to identify the regressor that best matches the inductive bias of the variable selection technique.*

Table 3 presents the *P*-values of the one-sided test of each method against each other method. A rough ranking of the four methods that select small variable sets, in terms of the predictive performance of the resulting models is SMMPC > BVS′ > LS > FS. Table 4 shows for each variable selection method the best match for regression method. In terms of CI, methods that select a high number of variables (NS and US) are better coupled with Ridge Cox Regression (recall also that AFT and Cox Regression are not applicable when the number of variables is higher than the number of non-censored samples). LS has a tendency to be better coupled with RSF, while for the remaining methods the best regressor is highly dependent on the dataset. In terms of IBS, Ridge Cox regression is the most frequently chosen method (Supplementary Table 8).

6.2 Comparing survival regression methods

In this section, we turn the focus on the survival regression methods. The nested cross-validation procedure is employed for parameter optimization and estimation of performance. CI performances are shown in Table 5 for the variable selection method that best matches the regression method (the *P*-values of the one-sided test comparing the performance of each method against each other method are give in the Supplementary Table 7). When considering the CI metric, a rough ranking of the methods is Ridge > SVCR > {RSF, AFT} > Cox. When we optimize for IBS, Ridge Cox Regression remains superior to all other methods and RSF significantly outperforms AFT models (Supplementary Tables 5 and 7). In terms

Table 5. Nested cross-validated performances of regression methods matched with the best-performing variable selection method

	AFT	Ridge	RSF	SVCR	Cox
Vijver	0.715 (0.03)	0.717 (0.02)	0.724 (0.02)	0.733 (0.03)	0.707 (0.03)
Veer	0.707 (0.1)	0.688 (0.08)	0.680 (0.08)	0.690 (0.08)	0.662 (0.05)
Ros.2002	0.625 (0.05)	0.631 (0.07)	0.622 (0.04)	0.629 (0.05)	0.638 (0.08)
Ros.2003	0.659 (0.09)	0.724 (0.12)	0.707 (0.52)	0.722 (0.20)	0.669 (0.07)
Bullinger	0.645 (0.09)	0.669 (0.08)	0.653 (0.07)	0.657 (0.08)	0.646 (0.07)
Beer	0.720 (0.17)	0.770 (0.16)	0.688 (0.08)	0.734 (0.13)	0.683 (0.15)
Mean CI	0.679 (0.09)	0.700 (0.09)	0.679 (0.14)	0.694 (0.09)	0.668 (0.07)
Mean IBS	0.176 (0.04)	0.162 (0.03)	0.166 (0.03)		0.174 (0.05)

The numbers in parentheses show the range of the metric between the best and the worst feature selection method tried. Ridge, Ridge regression; SVCR, Support Vector Censored Regression; Cox, Cox regression.

of CI, regressors that are only able to handle a relatively low number of variables, such as AFT and Cox are best-matched with SMMPC and BVS′. The rest are best-matched with US and BVS′. In term of IBS, there is no obvious pattern (see Supplementary Table 9).

7 DISCUSSION AND CONCLUSION

We adopt a successful variable selection method for classification for survival analysis tasks, named SMMPC. SMMPC is relatively simple, scalable to high-dimensional data and has a structural BN interpretation: it identifies the neighbors of T (the time-to-event) in the network representing the data distribution. In an extensive comparative evaluation over several datasets and regression methods, SMMPC is shown to statistically significantly outperform Bayesian Selection based on MCMC techniques, Bayesian Selection based on a greedy-search modified for high-dimensional data, Lasso Cox Regression (treated as a variable selection procedure), FS and US restricted to return as many variables as SMMPC. SMMPC selects fewer than a dozen variables per dataset making analysis of the models amenable to human experts. Relative to the other methods tried, SMMPC is 'stable' in terms of the returned variables (Supplementary Section 5). Finally, SMMPC shows a sharp learning curve in terms of available sample size (Supplementary Section 6). Further analyzing the results, shows that the performance of the regression methods is ordered as Ridge Cox Regression > SVCR > {RSFs and AFT} > Cox Regression. This conclusion is in agreement with prior studies pointing out the effectiveness and advantages of Ridge Cox Regression (Bovelstad *et al.*, 2007). SMMPC stems from the BN theory, causal induction theory and variable selection using the Markov blanket concept. The results show that these theories and corresponding variable selection algorithms may transfer their performance and scalability qualities to other tasks than classification. Thus, the extension of MMPC to different variable selection tasks (e.g. to time-series) seems promising.

Funding: STREP project 'HEARTFAID' (FP6-IST-2004-027107); European Research Consortium for Informatics and Mathematics (to V.L.).

Conflict of Interest: none declared.

REFERENCES

Aliferis,C.F. *et al.* (2003) HITON, a novel Markov blanket algorithm for optimal variable selection. In *Poceedings of the American Medical Informatics Association*, pp. 21–25.

Aliferis,C.F. *et al.* (2010) Local causal and Markov blanket induction algorithms for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. *J. Mach. Learn. Res.*, **11**, 171–234.

Bair,E. and Tibshirani,R (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.*, **2**, 511–522.

Beer,D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.

Bovelstad,H. *et al.* (2007) Predicting survival from microarray data a comparative study. *Bioinformatics*, **23**, 2080–2087.

Breiman,L. and Schapire,E. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Brown,L.E. and Tsamardinos,I. (2008) Markov blanket-based variable selection in feature space. *Technical Report DSL TR-08-01*. Department Biomedical Informatics, Vanderbilt University.

Bullinger,L. *et al.* (2004) Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N. Engl. J. Med.*, **350**, 1605–1616.

Cox,D.R. (1972) Regression models and life-tables. *J. R. Stat. Soc.*, **34**, 187–220.

Dudoit,S. and van der Laan,M. J. (2005) Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat. Methodol.*, **2**, 131–154.

Faraggi,D. and Simon,R. (1998) Bayesian variable selection method for censored survival data. *Biometrics*, **54**, 1475–1485.

Graf,E. *et al*. (1999) Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.*, **18**, 2529–2545.

Guyon,I. *et al*. (2004) Result analysis of the NIPS 2003 feature selection challenge. In *Advances in Neural Information Processing Systems 17*. MIT Press, Boston, MA, pp. 545–552.

Harrel,F.E. (2001) *Regression Modeling Strategies, With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, New York, NY.

Hastie,T. *et al*. (2001) Supervised harvesting of expression trees. *Genome Biol.*, **2**, research0003.1–research0003.12.

Hastings,W. K. (1970) Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Heagerty,P.J. *et al*. (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, **56**, 337–334.

Hoerl,A.E. and Kennard,R.W. (2000) Ridge Regression: biased estimation for nonorthogonal problems. *Technometrics*, **42**, 80–86.

Hothorn,T. *et al*. (2004) Bagging survival trees. *Stat. Med.*, **23**, 77–91.

Jensen,D. and Cohen,P. (2000) Multiple comparisons in induction algorithms. *Mach. Learn.*, **38**, pp. 309–338.

Klein,J.P. and Moeschberger,M.L. (2003) *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, New York, NY.

Kohavi,R. and John,G.H. (1997) Wrappers for feature subset selection. *Artif. Intell.*, **97**, 273–324.

Li,H. and Gui,J. (2004) Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, **20**, 208–215.

Nguyen,D.V. and Rocke,D.M. (2002) Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, **18**, 1625–1632.

Nguyen,T.S. and Rojo,J. (2009). Dimension reduction of microarray data in the presence of a censored survival response: a simulation study. *Stat. Appl. Genet. Mol. Biol.*, **8**, Article 4.

Pearl,J. (2000) *Causality, Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK.

Ripley,B.D. and Ripley,R.M. (1998) Neural networks as statistical methods in survival analysis. In *Artificial Neural Networks: Prospects for Medicine*. Landes Biosciences Publishers, Austin, TX, pp. 237–255.

Rosenwald,A. *et al.* (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. *N. Engl. J. Med.*, **346**, 1937–1947.

Rosenwald,A. *et al.* (2003) The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell*, **3**, 185–197.

Sha,N. *et al.* (2006) Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics*, **22**, 2262–2268.

Shivaswamy,P.K. *et al.* (2007) A support vector approach to censored targets. In *ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*. IEEE Computer Society, Washington, DC, USA, pp. 655–660.

Sohn,I. *et al*. (2009) Gradient lasso for Cox proportional hazards model. *Bioinformatics*, **25**, 1775–1781.

Spirtes,P. *et al*. (2000) *Causation, Prediction, and Search*. 2nd edn.MIT Press, Cambridge, MA.

Statnikov,A. *et al*. (2005) GEMS: a system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. *Int. J. Med. Inform.*, **74**, 491–503.

Tibshirani,R. (1997a) The lasso method for variable selection in the Cox model. *Stat. Med.*, **16**, 385–395.

Tibshirani,R. (1997b) The lasso method for variable selection in the Cox model. *Stat. Med.*, **16**, 385–395.

Tsamardinos,I. and Aliferis,C.F. (2003) Towards principled feature selection: relevancy, filters and wrappers. In *Ninth International Workshop on Artificial Intelligence and Statistics 2003*. Morgan Kaufmann Publishers, San Francisco, CA, USA.

Tsamardinos,I. and Brown,L. E. (2008) Bounding the false discovery rate in local bayesian network learning. In *AAAI'08: Proceedings of the 23rd National Conference on Artificial Intelligence*. AAAI Press, Menlo Park, California, pp. 1100–1105.

Tsamardinos,I. *et al*. (2003) Time and sample efficient discovery of Markov blankets and direct causal relations. In *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 673–678.

Tsamardinos,I. *et al*. (2006) The Max–Min Hill-Climbing Bayesian network structure learning algorithm. *Mach. Learn.*, **65**, 31–78.

van't Veer,L. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

van de Vijver,M. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2000.

van Wieringen,W.N. *et al*. (2009) Survival prediction using gene expression data: a review and comparison. *Comput. Stat. Data Anal.*, **53**, 1590–1603.

Witten,D.M. and Tibshirani,R. (2009) Survival analysis with high-dimensional covariates. *Stat. Methods Med. Res.*, **19**, 29–51.