

RANK DISCRIMINANTS FOR PREDICTING PHENOTYPES FROM RNA EXPRESSION

BY BAHMAN AFSARI[‡], ULISSES BRAGA NETO[§] AND DONALD GEMAN[‡]

Johns Hopkins University[‡] and Texas A&M University[§]

Statistical methods for analyzing large-scale biomolecular data are commonplace in computational biology. A notable example is phenotype prediction from gene expression data, for instance detecting human cancers, differentiating subtypes, and predicting clinical outcomes. Still, clinical applications remain scarce. One reason is that the complexity of the decision rules that emerge from standard statistical learning impedes biological understanding, in particular any mechanistic interpretation. Here we explore decision rules for binary classification utilizing only the ordering of expression among several genes; the basic building blocks are then two-gene expression comparisons. The simplest example, just one comparison, is the *TSP* classifier, which has appeared in a variety of cancer-related discovery studies. Decision rules based on multiple comparisons can better accommodate class heterogeneity, and thereby increase accuracy, and might provide a link with biological mechanism. We consider a general framework (“rank-in-context”) for designing discriminant functions, including a data-driven selection of the number and identity of the genes in the support (“context”). We then specialize to two examples: voting among several pairs and comparing the median expression in two groups of genes. Comprehensive experiments assess accuracy relative to other, more complex, methods, and reinforce earlier observations that simple classifiers are competitive.

1. Introduction. Statistical methods for analyzing high-dimensional biomolecular data generated with high-throughput technologies permeate the literature in computational biology. Such analyses have uncovered a great deal of information about biological processes, such as important mutations and lists of “marker genes” associated with common diseases (Jones et al. (2008); Thomas et al. (2007)) and key interactions in transcriptional regulation (Auffray (2007); Lee et al. (2008)). Our interest here is learning classifiers that can distinguish between cellular phenotypes from mRNA transcript levels collected from cells in assayed tissue, with a primary focus on the structure of the prediction rules. Our work is motivated by applica-

[‡]Supported by NIH-NCRR Grant UL1 RR 025005.

[§]Supported by the National Science Foundation through NSF award CCF-0845407.

Keywords and phrases: Cancer classification, Gene expression, Rank discriminant, Order statistics

TABLE 1

The Datasets: Twenty-one datasets involving two disease-related phenotypes (e.g., cancer vs normal tissue or two cancer sub-types), illustrating the “small n, large d” situation. The more pathological phenotype is labeled as Class 1 when this information is available.

	Study	Class 0 (size)	Class 1 (size)	Probes d	Reference
D1	Colon	Normal (22)	Tumor (40)	2000	Alon et al. (1999)
D2	BRCA1	non-BRCA1 (93)	BRCA1 (25)	1658	Lin et al. (2009)
D3	CNS	Classic (25)	Desmoplastic (9)	7129	Pomeroy et al. (2002)
D4	DLBCL	DLBCL (58)	FL (19)	7129	Shipp et al. (2002)
D5	Lung	Mesothelioma (150)	ADCS (31)	12533	Gordon et al. (2002)
D6	Marfan	Normal (41)	Marfan (60)	4123	Yao et al. (2007)
D7	Crohn’s	Normal (42)	Crohn’s (59)	22283	Burczynski et al. (2006)
D8	Sarcoma	GIST (37)	LMS (31)	43931	Price et al. (2007)
D9	Squamous	Normal (22)	Head-Neck (22)	12625	Kuriakose et al. (2004)
D10	GCM	Normal (90)	Tumor (190)	16063	Ramaswamy et al. (2001)
D11	Leukemia 1	ALL (25)	AML (47)	7129	Golub et al. (1999)
D12	Leukemia 2	AML1 (24)	AML2 (24)	12564	Armstrong et al. (2002)
D13	Leukemia 3	ALL(710)	AML (501)	19896	Kohlmann et al. (2008)
D14	Leukemia 4	Normal (138)	AML (403)	19896	Mills et al. (2009)
D15	Prostate 1	Normal (50)	Tumor (52)	12600	Singh et al. (2002)
D16	Prostate 2	Normal (38)	Tumor (50)	12625	Stuart et al. (2004)
D17	Prostate 3	Normal (9)	Tumor (24)	12626	Welsh et al. (2001)
D18	Prostate 4	Normal (25)	Primary (65)	12619	Yao et al. (2004)
D19	Prostate 5	Primary (25)	Metastatic (65)	12558	Yao et al. (2004)
D20	Breast 1	ER-positive (61)	ER-negative(36)	16278	Enerly et al. (2011)
D21	Breast 2	ER-positive(127)	ER-negative(80)	9760	Buffa et al. (2011)

tions to genetic diseases such as cancer, where malignant phenotypes arise from the net effect of interactions among multiple genes and other molecular agents within biological networks. Statistical methods can enhance our understanding by detecting the presence of disease (e.g., “tumor” vs “normal”), discriminating among cancer sub-types (e.g., “GIST” vs “LMS” or “BRCA1 mutation” vs “no BRCA1 mutation”) and predicting clinical outcomes (e.g., “poor prognosis” vs “good prognosis”).

Whereas the need for statistical methods in biomedicine continues to grow, the effects on clinical practice of existing classifiers based on gene expression are widely acknowledged to remain limited; see Altman et al. (2011), Marshall (2011), Evans et al. (2011) and the discussion in Winslow et al. (2012). One barrier is the study-to-study diversity in reported prediction accuracies and “signatures” (lists of discriminating genes). Some of this variation can be attributed to the over-fitting that results from the unfavorable ratio of the sample size to the number of potential biomarkers, i.e., the infamous “small n, large d” dilemma. Typically, the number of samples (chips, profiles, patients) per class is $n = 10 - 1000$ whereas the number of features (exons, transcripts, genes) is usually $d = 1000 - 50,000$; Table 1 displays the sample sizes and the numbers of features for twenty-one publicly available datasets involving two phenotypes.

Complex decision rules are obstacles to mature applications. The classification methods applied to biological data were usually designed for other purposes, such as improving statistical learning or applications to vision and speech, with little emphasis on transparency. Specifically, the rules gen-

erated by nearly all standard, off-the-shelf techniques applied to genomics data, such as neural networks (Bicciato et al. (2003), Bloom et al. (2004), Khan et al. (2001)), multiple decision trees (Boulesteix, Tutz and Strimmer (2003), Zhang, Yu and Singer (2003)), support vector machines (Peng et al. (2003), Yeang et al. (2001)), boosting (Qu et al. (2002), Dettling and Buhlmann (2003)), and linear discriminant analysis (Guo, Hastie and Tibshirani (2007), Tibshirani et al. (2002)), usually involve nonlinear functions of hundreds or thousands of genes, a great many parameters, and are therefore too complex to characterize mechanistically.

In contrast, follow-up studies, for instance independent validation or therapeutic development, are usually based on a relatively small number of biomarkers and usually require an understanding of the role of the genes and gene products in the context of molecular pathways. Ideally, the decision rules could be interpreted mechanistically, for instance in terms of transcriptional regulation, and be robust with respect to parameter settings. Consequently, what is notably missing from the large body of work applying classification methodology to computational genomics is a solid link with potential mechanisms, which seem to be a necessary condition for “translational medicine” (Winslow et al. (2012)) i.e., drug development and clinical diagnosis.

These translational objectives, and small-sample issues, argue for limiting the number of parameters and introducing strong constraints. The two principal objectives for the family of classifiers described here are:

- Use elementary and parameter-free building blocks to assemble a classifier which is determined by its support.
- Demonstrate that such classifiers can be as discriminating as those that emerge from the most powerful methods in statistical learning.

The building blocks we choose are two-gene comparisons, which we view as “biological switches” which can be directly related to regulatory “motifs” or other properties of transcriptional networks. The decision rules are then determined by expression orderings. However, explicitly connecting statistical classification and molecular mechanism for particular diseases is a major challenge and is well beyond the scope of this paper; by our construction we are anticipating our longer-term goal of incorporating mechanism by delineating candidate motifs using prior biological knowledge. Some comments on the relationship between comparisons and regulation appear in the concluding section.

To meet our second objective, we measure the performance of our comparison-based classifiers relative to two popular alternatives, namely support vector

machines and *PAM* (Tibshirani et al. (2002)), a variant of linear discriminant analysis. The “metric” chosen is the estimated error in multiple runs of ten-fold cross validation for each of the twenty-one real datasets in Table 1. (Computational cost is not considered; applying any of our comparison-based decision rules to a new sample is virtually instantaneous.) Whereas a comprehensive simulation study could be conducted, for example along the lines of those in Guo, Hastie and Tibshirani (2005), Zhang et al. (2006) and Fan and Fan (2008) based on Gaussian models of microarray data, rather our intention is different: show that even when the number of parameters is small, in fact the decision rule is determined by the support, the accuracy measured by cross-validation on real data is no worse than with currently available classifiers.

More precisely, all the classifiers studied in this paper are based on a general *rank discriminant* $g(\mathbf{X}; \Theta)$, a real-valued function on the ranks of \mathbf{X} over a (possibly ordered) subset of genes Θ , called the *context* of the classifier. We are searching for characteristic perturbations in this ordering from one phenotype to another. The *TSP* classifier is the simplest example (see Section 2), and the decision rule is illustrated in Figure 1. This data set has expression profiles for two kinds of gastrointestinal cancer (gastrointestinal stromal-GIST, leiomyosarcoma-LMS) which are difficult to distinguish clinically but require very different treatments (Price et al. (2007)). Each point on the x-axis corresponds to a sample, and the vertical dashed line separates the two phenotypes. The y-axis represents expression; as seen, the “reversal” of the ordering of the expressions of the two genes identifies the phenotype except in two samples.

Evidently, a great deal of information may be lost by converting to ranks, particularly if the expression values are high resolution. But there are technical advantages to basing prediction on ranks, including reducing study-to-study variations due to data normalization and pre-processing. Rank-based methods are evidently invariant to general monotone transformations of the original expression values, such as the widely-used quantile normalization (Bloomfield, Irizarry and Speed (2004)). Thus, methods based on ranks can combine inter-study microarray data without the need to perform data normalization, thereby increasing sample size.

However, our principal motivation is complexity reduction: severely limiting the number of variables and parameters, and in fact introducing what we call *rank-in-context* (RIC) discriminants which depend on the training data only through the context. The classifier f is then defined by thresholding g . This implies that, given a context Θ , the RIC classifier corresponds to a *fixed* decision boundary, in the sense that it does not depend on the train-

ing data. This sufficiency property helps to reduce variance by rendering the classifiers relatively insensitive to small disturbances to the ranks of the training data and is therefore especially suitable to small-sample settings. Naturally, the performance critically depends on the appropriate choice of Θ . We propose a simple yet powerful procedure to select Θ from the training data, partly inspired by the principle of analysis of variance and involving the sample means and sample variances of the empirical distribution of g under the two classes. In particular, we do not base the choice directly on minimizing error.

We consider two examples of the general framework. The first is a new method for learning the context of *KTSP*, a previous extension of *TSP* to a variable number of pairs. The decision rule of the *KTSP* classifier is the majority vote among the top k pairs of genes, illustrated in Figure 1 for $k = 10$ for the same dataset as above. In previous statistical and applied work (Tan et al. (2005)), the parameter K (the number of comparisons) was determined by an inner loop of cross-validation, which is subject to over-fitting with small samples. We also propose comparing the median of expression between two sets of genes; this *Top-Scoring Median (TSM)* rule is also illustrated in Figure 1; as can be seen, the difference of the medians generally has a larger “margin” than in the special case of singleton sets, i.e., *TSP*. A summary of all the methods is given in Table 2.

After reviewing related work in the following section, in Section 3 we present the classification scenario, propose our general statistical framework, and focus on two examples: *KTSP* and *TSM*. The experimental results are in Section 4, where comparisons are drawn, and we conclude with some discussion about the underlying biology in Section 5.

2. Previous and Related Work. Our work builds on previous studies analyzing transcriptomic data solely based on the *relative expression* among a small number of transcripts. The simplest example, the Top-Scoring Pair (*TSP*) classifier, was introduced in Geman et al. (2004) and is based on two genes. Various extensions and illustrations appeared in Xu et al. (2005), Lin et al. (2009), Tan et al. (2005). Applications to phenotype classification include differentiating between stomach cancers (Price et al. (2007)), predicting treatment response in breast cancer (Weichselbaum et al. (2008)) and acute myeloid leukemia (Raponi et al. (2008)), detecting BRCA1 mutations (Lin et al. (2009)), grading prostate cancers (Zhao, Logothetis and Gorlov (2010)), and separating diverse human pathologies assayed through blood-borne leukocytes (Edelman et al. (2009)).

In Geman et al. (2004) and subsequent papers about *TSP*, the discrimi-

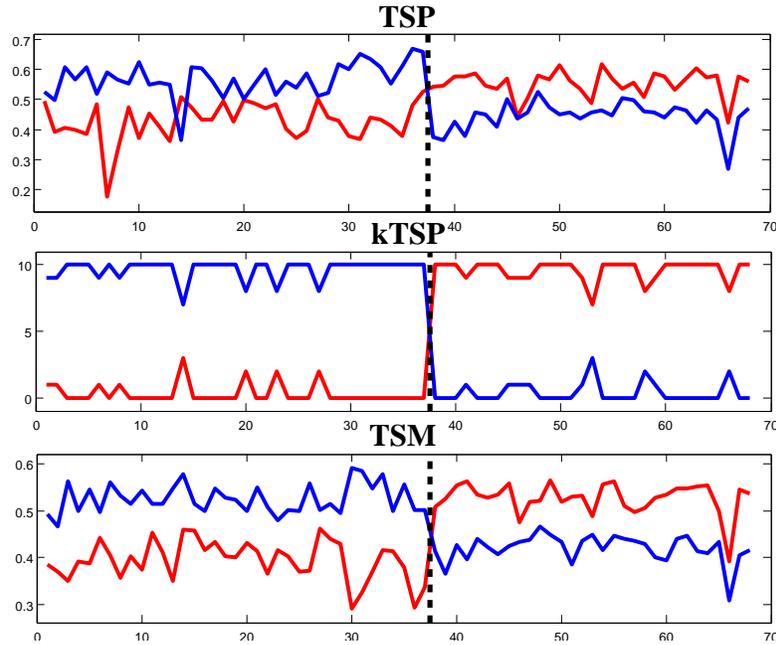


FIG 1. Results of three rank-based classifiers for differentiating two cancer subtypes, *GIST* and *LMS*. The training set consists of 37 *GIST* samples and 31 *LMS* samples (separated by the vertical dashed line); each sample provides measurements for 43,931 transcripts. **TSP**: Expression values for the two genes selected by the TSP algorithm. **kTSP**: The number of votes for each class among the $K = 10$ pairs of genes selected by KTSP algorithm. **TSM**: Median expressions of two sets of genes selected by the TSM algorithm.

nating power of each pair of genes i, j was measured by the absolute difference between the probabilities of the event that gene i is expressed more than gene j in the two classes. These probabilities were estimated from training data and (binary) classification resulted from voting among all top-scoring pairs. In Xu et al. (2005) a secondary score was introduced which provides a *unique* top-scoring pair. In addition, voting was extended to the k highest-scoring pairs of genes. The motivation for this *KTSP* classifier and other extensions (Tan et al. (2005), Anderson et al. (2007), Xu, Geman and Winslow (2007)) is that more genes may be needed to detect cancer pathogenesis, especially if the principle objective is to characterize as well as recognize the process. Finally, in a precursor to the work here (Xu, Geman and Winslow (2007)), the two genes in *TSP* were replaced by two equally-sized *sets* of genes and the average ranks were compared. Since the direct extension of *TSP* score maximization was computationally impossible, and likely to badly over-fit the data, the sets were selected by splitting top-scoring pairs and

repeated random sampling. Although ad hoc, this process further demonstrated the discriminating power of rank statistics for microarray data.

Finally, there is some related work about ratios of concentrations (which are natural in chemical terms) for diagnosis and prognosis. That work is not rank-based but retains invariance to scaling. Golub et al. (1999) distinguished between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung by combining multiple ratios into a single diagnostic tool, and Ma et al. (2004) found that a two-gene expression ratio derived from a genome-wide, oligonucleotide microarray analysis of estrogen receptor (ER)-positive, invasive breast cancers predicts tumor relapse and survival in patients treated with tamoxifen, which is crucial for early-stage breast cancer management.

3. Rank-In-Context Classification. In this section, we introduce a general framework for rank-based classifiers using comparisons among a limited number of gene expressions, called the *context*. In addition, we describe a general method to select the context, which is inspired by the analysis of variance paradigm of classical statistics. These classifiers have the RIC property that they depend on the sample training data solely through the context selection; in other words, given the context, the classifiers have a fixed decision boundary and do not depend on any further learning from the training data. For example, as will be seen in later sections, the *Top-Scoring Pair (TSP)* classifier is RIC. Once a pair of genes (i.e., the context) is specified, the *TSP* decision boundary is fixed, and corresponds to a 45-degree line going through the origin in the feature space defined by the two genes. This property confers to RIC classifiers a *minimal-training* property, which makes them insensitive to small disturbances to the ranks of the training data, reducing variance and overfitting, and rendering them especially suitable to the $n \ll d$ settings illustrated in Table 1. We will demonstrate the general RIC framework with two specific examples, namely the previously introduced *KTSP* classifier based on majority voting among comparisons (Tan et al. (2005)), as well as a new classifier based on the comparison of the medians, the *Top-Scoring Medians (TSM)* classifier.

3.1. RIC Discriminant. Let $\mathbf{X} = (X_1, X_2, \dots, X_d)$ denote the expression values of d genes on an expression microarray. Our objective is to use \mathbf{X} to distinguish between two conditions or phenotypes for the cells in the assayed tissue, denoted $Y = 0$ and $Y = 1$. A classifier f associates a label $f(\mathbf{X}) \in \{0, 1\}$ with a given expression vector \mathbf{X} . Practical classifiers are inferred from training data, consisting of i.i.d. pairs $S_n = \{(\mathbf{X}^{(1)}, Y^{(1)}), \dots, (\mathbf{X}^{(n)}, Y^{(n)})\}$.

The classifiers we consider in this paper are all defined in terms of a general *rank-in-context discriminant* $g(\mathbf{X}; \Theta(S_n))$, which is defined as a real-valued function on the ranks of \mathbf{X} over a subset of genes $\Theta(S_n) \subset \{1, \dots, d\}$, which is determined by the training data S_n and is called the *context* of the classifier (the order of indices in the context may matter). The corresponding *RIC classifier* f is defined by

$$(1) \quad f(\mathbf{X}; \Theta(S_n)) = \mathbf{I}(g(\mathbf{X}; \Theta(S_n)) > t) = \begin{cases} 1, & g(\mathbf{X}; \Theta(S_n)) > t \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathbf{I}(E)$ denotes the indicator variable of event E . The threshold parameter t can be adjusted to achieve a desired specificity and sensitivity (see Section 3.4 below); otherwise, one usually sets $t = 0$. For simplicity we will write Θ instead of $\Theta(S_n)$, with the implicit understanding that in RIC classification Θ is selected from the training data S_n .

We will consider two families of RIC classifiers. The first example is the *k-Top Scoring Pairs (KTSP)* classifier, which is a majority voting rule among k pairs of genes (Tan et al. (2005)); *KTSP* was the winning entry of the International Conference in Machine Learning and Applications (ICMLA) 2008 challenge for micro-array classification (Geman et al. (2008)). Here, the context is partitioned into a set of gene pairs $\Theta = \{(i_1, j_1), \dots, (i_k, j_k)\}$, where k is a positive odd integer, in such a way that all pairs are disjoint, i.e., all $2k$ genes are distinct. The RIC discriminant is given by:

$$(2) \quad g_{\text{KTSP}}(\mathbf{X}; (i_1, j_1), \dots, (i_k, j_k)) = \sum_{r=1}^k \left[\mathbf{I}(X_{i_r} < X_{j_r}) - \frac{1}{2} \right].$$

This *KTSP* RIC discriminant simply counts positive and negative “votes” in favor of ascending or descending ranks, respectively. The *KTSP* classifier is given by (1), with $t = 0$, which yields

$$(3) \quad f_{\text{KTSP}}(\mathbf{X}; (i_1, j_1), \dots, (i_k, j_k)) = \mathbf{I} \left(\sum_{r=1}^k \mathbf{I}(X_{i_r} < X_{j_r}) > \frac{k}{2} \right).$$

The *KTSP* classifier is thus a majority-voting rule: it assigns label 1 to the expression profile if the number of ascending ranks exceeds the number of descending ranks in the context. The choice of odd k avoids the possibility of a tie in the vote. If $k = 1$, then the *KTSP* classifier reduces to $f_{\text{TSP}}(\mathbf{X}; (i, j)) = \mathbf{I}(X_i < X_j)$, the *Top-Scoring Pair (TSP)* classifier (Geman et al. (2004)).

The second example of an RIC classifier we propose is the *Top Scoring Median (TSM)* classifier, which compares the median rank of two sets of genes. The median rank has the advantage that for any individual sample the median is the value of one of the genes. Hence, in this sense, a comparison of medians for a given sample is equivalent to the comparison of two gene expressions, as in the *TSP* decision rule. Here, the context is partitioned into two sets of genes, $\Theta = \{G_k^+, G_k^-\}$, such that $|G_k^+| = |G_k^-| = k$, where k is again a positive odd integer, and G_k^+ and G_k^- are disjoint, i.e., all $2k$ genes are distinct. Let R_i be the rank of X_i in the context $\Theta = G_k^+ \cup G_k^-$, such that $R_i = j$ if X_i is the j th smallest value among the gene expression values indexed by Θ . The RIC discriminant is given by:

$$(4) \quad g_{\text{TSM}}(\mathbf{X}; G_k^+, G_k^-) = \text{med}_{j \in G_k^+} R_j - \text{med}_{i \in G_k^-} R_i.$$

where “med” denotes the median operator. The *TSM* classifier is then given by (1), with $t = 0$, which yields

$$(5) \quad f_{\text{TSM}}(\mathbf{X}; G_k^+, G_k^-) = \mathbf{I} \left(\text{med}_{j \in G_k^+} R_j > \text{med}_{i \in G_k^-} R_i \right).$$

Therefore, the *TSM* classifier outputs 1 if the median of ranks in G_k^+ exceeds the median of ranks in G_k^- , and 0 otherwise. Notice that this is equivalent to comparing the medians of the raw expression values directly. We remark that an obvious variation would be to compare the average rank rather than the median rank, which corresponds to the “TSPG” approach defined in Xu, Geman and Winslow (2007), except that in that study, the context for TSPG was selected by splitting a fixed number of TSPs. We observed that the performances of the mean-rank and median-rank classifiers are similar, with a slight superiority of the median-rank (data not shown).

3.2. Criterion for Context Selection. The performance of RIC classifiers critically depends on the appropriate choice of the context $\Theta \subset \{1, \dots, d\}$. We propose a simple yet powerful procedure to select Θ from the training data S_n . To motivate the proposed criterion, first note that a necessary condition for the context Θ to yield a good classifier is that the discriminant $g(\mathbf{X}; \Theta)$ has sufficiently distinct distributions under $Y = 1$ and $Y = 0$. This can be expressed by requiring that the difference between the expected values of $g(\mathbf{X}; \Theta)$ *between* the populations, namely

$$(6) \quad \delta(\Theta) = E[g(\mathbf{X}; \Theta) | Y = 1, S_n] - E[g(\mathbf{X}; \Theta) | Y = 0, S_n]$$

be maximized. Notice that this maximization is with respect to Θ alone; g is fixed and chosen *a priori*. In practice, one employs the maximum-likelihood empirical criterion

$$(7) \quad \hat{\delta}(\Theta) = \hat{E}[g(\mathbf{X}; \Theta) \mid Y = 1, S_n] - \hat{E}[g(\mathbf{X}; \Theta) \mid Y = 0, S_n],$$

where

$$(8) \quad \hat{E}[g(\mathbf{X}; \Theta) \mid Y = c, S_n] = \frac{\sum_{i=1}^n g(\mathbf{X}^{(i)}; \Theta) \mathbf{I}(Y^{(i)} = c)}{\sum_{i=1}^n \mathbf{I}(Y^{(i)} = c)},$$

for $c = 0, 1$.

In the case of *KTSP*, the criterion in (6) becomes

$$(9) \quad \delta_{\text{KTSP}}((i_1, j_1), \dots, (i_k, j_k)) = \sum_{r=1}^k s_{i_r j_r}$$

where the *pairwise score* s_{ij} for the pair of genes (i, j) is defined as

$$(10) \quad s_{ij} = P(X_i < X_j \mid Y = 1) - P(X_i < X_j \mid Y = 0).$$

Notice that if the pair of random variables (X_i, X_j) has a continuous distribution, so that $P(X_i = X_j) = 0$, then $s_{ij} = -s_{ji}$. In this case $X_i < X_j$ can be replaced by $X_i \leq X_j$ in s_{ij} in (10).

The empirical criterion $\hat{\delta}_{\text{KTSP}}((i_1, j_1), \dots, (i_k, j_k))$ (c.f. Eq 7) is obtained by substituting in (9) the *empirical pairwise scores*

$$(11) \quad \hat{s}_{ij} = \hat{P}(X_i < X_j \mid Y = 1) - \hat{P}(X_i < X_j \mid Y = 0).$$

Here the empirical probabilities are defined by $\hat{P}(X_i < X_j \mid Y = c) = \hat{E}[\mathbf{I}(X_i < X_j) \mid Y = c]$, for $c = 0, 1$, where the operator \hat{E} is defined in (8).

For *TSM*, the criterion in (6) is given by

$$(12) \quad \delta_{\text{TSM}}(G_k^+, G_k^-) = E \left[\text{med}_{j \in G_k^+} R_j - \text{med}_{i \in G_k^-} R_i \mid Y = 1 \right] - E \left[\text{med}_{j \in G_k^+} R_j - \text{med}_{i \in G_k^-} R_i \mid Y = 0 \right].$$

Proposition S1 in Supplement A (Afsari, Braga-Neto and Geman (2014a)) shows that, under some assumptions,

$$(13) \quad \delta_{\text{TSM}}(G_k^+, G_k^-) = \frac{2}{k} \sum_{i \in G_k^-, j \in G_k^+} s_{ij},$$

where s_{ij} is defined in (10).

The difference between the two criteria (9) for *KTSP* and (13) for *TSM* for selecting the context is that the former involves scores for k expression comparisons and the latter involves k^2 comparisons since each gene $i \in G_k^-$ is paired with each gene $j \in G_k^+$. Moreover, using the estimated solution to maximizing (9) (see below) to construct G_k^- and G_k^+ by putting the first gene from each pair into one and the second gene from each pair into the other does not work as well in maximizing (13) as the algorithms described below.

The distributional smoothness conditions Proposition S1 are justified if k is not too large (see Supplement A). Finally, the empirical criterion $\hat{\delta}_{\text{TSM}}(G_k^+, G_k^-)$ can be calculated by substituting in (13) the *empirical pairwise scores* \hat{s}_{ij} defined in (11).

3.3. Maximization of the Criterion. Maximization of (6) or (7) works well as long as the *size* of the context $|\Theta|$, i.e., the number of context genes, is kept fixed, because the criterion tends to be monotonically increasing with $|\Theta|$, which complicates selection. We address this problem by proposing a modified criterion, which is partly inspired by the principle of analysis of variance in classical statistics. This modified criterion penalizes the addition of more genes to the context by requiring that the variance of $g(\mathbf{X}; \Theta)$ *within* the populations be minimized. The latter is given by

$$(14) \quad \hat{\sigma}(\Theta) = \sqrt{\widehat{\text{Var}}(g(\mathbf{X}; \Theta) \mid Y = 0, S_n) + \widehat{\text{Var}}(g(\mathbf{X}; \Theta) \mid Y = 1, S_n)},$$

where $\widehat{\text{Var}}$ is the maximum-likelihood estimator of the variance,

$$\begin{aligned} & \widehat{\text{Var}}(g(\mathbf{X}; \Theta) \mid Y = c, S_n) \\ &= \frac{\sum_{i=1}^n (g(\mathbf{X}^{(i)}; \Theta) - \hat{E}[g(\mathbf{X}; \Theta) \mid Y = c, S_n])^2 \mathbf{I}(Y^{(i)} = c)}{\sum_{i=1}^n \mathbf{I}(Y^{(i)} = c)}, \end{aligned}$$

for $c = 0, 1$. The modified criterion to be maximized is

$$(15) \quad \hat{\tau}(\Theta) = \frac{\hat{\delta}(\Theta)}{\hat{\sigma}(\Theta)},$$

The statistic $\hat{\tau}(\Theta)$ resembles the Welch two-sample t-test statistic of classical hypothesis testing (Casella and Berger (2002)).

Direct maximization of (7) or (15) is in general a hard computational problem for the numbers of genes typically encountered in expression data. We propose instead a greedy procedure. Assuming that a pre-defined range of values Ω for the context size $|\Theta|$ is given, the procedure is:

- (1) For each value of $k \in \Omega$, an optimal context Θ_k^* is chosen that maximizes (7) among all contexts Θ_k containing k genes:

$$\Theta_k^* = \arg \max_{|\Theta|=k} \hat{\delta}(\Theta).$$

- (2) An optimal value k^* is chosen that maximizes (15) among all contexts $\{\Theta_k^* \mid k \in \Omega\}$ obtained in the previous step:

$$k^* = \arg \max_{k \in \Omega} \hat{\tau}(\Theta_k^*).$$

For *KTSP*, the maximization in step (1) of the previous context selection procedure becomes

$$\begin{aligned} (16) \quad \{(i_1^*, j_1^*), \dots, (i_k^*, j_k^*)\} &= \arg \max_{\{(i_1, j_1), \dots, (i_k, j_k)\}} \hat{\delta}_{\text{KTSP}}((i_1, j_1), \dots, (i_k, j_k)) \\ &= \arg \max_{\{(i_1, j_1), \dots, (i_k, j_k)\}} \sum_{r=1}^k \hat{s}_{i_r j_r}. \end{aligned}$$

We propose a greedy approach to this maximization problem: initialize the list with the top scoring pair of genes, then keep adding pairs to the list whose genes have not appeared so far (ties are broken by the secondary score proposed in Xu et al. (2005)). This process is repeated until k pairs are chosen and corresponds essentially to the same method that was proposed, for fixed k , in the original paper on *KTSP* (Tan et al. (2005)). Thus, the previously-proposed heuristic has a justification in terms of maximizing the separation between the rank discriminant (2) across the classes.

To obtain the optimal value k^* , one applies step (2) of the context selection procedure, with a range of values $k \in \Omega = \{3, 5, \dots, K\}$, for odd K ($k = 1$ can be added if 1-TSP is considered). Note that here

$$(17) \quad \hat{\sigma}_{\text{KTSP}}(\Theta) = \sqrt{\widehat{\text{Var}} \left(\sum_{r=1}^k [\mathbf{I}(X_{i_r^*} < X_{j_r^*})] \mid Y=0 \right) + \widehat{\text{Var}} \left(\sum_{r=1}^k [\mathbf{I}(X_{i_r^*} < X_{j_r^*})] \mid Y=1 \right)}.$$

Therefore, the optimal value of k is selected by

$$(18) \quad k^* = \arg \max_{k=3,5,\dots,K} \hat{\tau}_{\text{KTSP}}((i_1^*, j_1^*), \dots, (i_k^*, j_k^*))$$

where

$$(19) \quad \begin{aligned} \hat{\tau}_{\text{KTSP}}((i_1^*, j_1^*), \dots, (i_k^*, j_k^*)) &= \frac{\hat{\delta}_{\text{KTSP}}((i_1^*, j_1^*), \dots, (i_k^*, j_k^*))}{\hat{\sigma}_{\text{KTSP}}((i_1^*, j_1^*), \dots, (i_k^*, j_k^*))} \\ &= \frac{\sum_{r=1}^k \hat{s}_{i_r^* j_r^*}}{\sqrt{\widehat{\text{Var}}\left(\sum_{r=1}^k [\mathbf{I}(X_{i_r^*} < X_{j_r^*})] \mid Y=0\right) + \widehat{\text{Var}}\left(\sum_{r=1}^k [\mathbf{I}(X_{i_r^*} < X_{j_r^*})] \mid Y=1\right)}}. \end{aligned}$$

Finally, the optimal context is then given by $\Theta^* = \{(i_1^*, j_1^*), \dots, (i_k^*, j_k^*)\}$.

For *TSM*, The maximization in step (1) of the context selection procedure can be written as

$$(20) \quad \begin{aligned} (G_k^{+,*}, G_k^{-,*}) &= \arg \max_{(G_k^+, G_k^-)} \hat{\delta}_{\text{TSM}}(G_k^+, G_k^-) \\ &= \arg \max_{(G_k^+, G_k^-)} \sum_{i \in G_k^-, j \in G_k^+} \hat{s}_{ij}, \end{aligned}$$

Finding the global maximum in (20) is not feasible in general. We consider a sub-optimal strategy for accomplishing this task: sequentially construct the context by adding two genes at a time. Start by selecting the *TSP* pair i, j and setting $G_1^- = \{i\}$ and $G_1^+ = \{j\}$. Then select the pair of genes i', j' distinct from i, j such that the sum of scores is maximized by $G_2^- = \{i, i'\}$ and $G_2^+ = \{j, j'\}$, i.e., $\hat{\delta}_{\text{TSM}}(G_k^+, G_k^-)$ is maximized over all sets G_k^+, G_k^- of size two, assuming $i \in G_k^-$ and $j \in G_k^+$. This involves computing three new scores. Proceed in this way until k pairs have been selected.

To obtain the optimal value k^* , one applies step (2) of the context selection procedure, with a range of values $k \in \Omega = \{3, 5, \dots, K\}$, for odd K (the choice of Ω is dictated by the facts that $k = 1$ reduces to 1-TSP, whereas Proposition S1 does not hold for even k):

$$k^* = \arg \max_{k=3,5,\dots,K} \hat{\tau}_{\text{TSM}}(G_k^{+,*}, G_k^{-,*})$$

where

$$\begin{aligned}
 (21) \quad \hat{\tau}_{\text{TSM}}(G_k^{+,*}, G_k^{-,*}) &= \frac{\hat{\delta}_{\text{TSM}}(G_k^{+,*}, G_k^{-,*})}{\hat{\sigma}_{\text{TSM}}(G_k^{+,*}, G_k^{-,*})} \\
 &= \frac{\hat{E} \left[\text{med}_{j \in G_k^{+,*}} R_j - \text{med}_{i \in G_k^{-,*}} R_i \mid Y = 1 \right] - \hat{E} \left[\text{med}_{j \in G_k^{+,*}} R_j - \text{med}_{i \in G_k^{-,*}} R_i \mid Y = 0 \right]}{\sqrt{\widehat{\text{Var}} \left(\text{med}_{j \in G_k^{+,*}} R_j - \text{med}_{i \in G_k^{-,*}} R_i \mid Y = 0 \right) + \widehat{\text{Var}} \left(\text{med}_{j \in G_k^{+,*}} R_j - \text{med}_{i \in G_k^{-,*}} R_i \mid Y = 1 \right)}}.
 \end{aligned}$$

Notice that $\hat{\tau}_{\text{TSM}}$ is defined directly by replacing (4) into (7) and (14), and then using (15). In particular, it does not use the approximation in (13). Finally, the optimal context is given by $\Theta^* = (G_{k^*}^{+,*}, G_{k^*}^{-,*})$.

For both *KTSP* and *TSM* classifiers, the step-wise process to perform the maximization of the criterion, c.f. Eqs. (16) and (20), does not need to be restarted as k increases, since the sub-optimal contexts are nested (by contrast, the method in Tan et al. (2005) employed cross-validation to choose k^*). The detailed context selection procedure for *KTSP* and *TSM* classifiers is given in Algorithms S1 and S2 in Supplement C (Afsari, Braga-Neto and Geman (2014c)).

3.4. Error Rates. In this section, we discuss the choice of the threshold t used in (1). The *sensitivity* is defined as $P(f(\mathbf{X}) = 1 \mid Y = 1)$ and the *specificity* is defined as $P(f(\mathbf{X}) = 0 \mid Y = 0)$. We are interested in controlling both, but trade-offs are inevitable. The choice of which phenotype to designate as 1 is application-dependent; often sensitivity is relative to the more malignant one and this is the way we have assigned labels to the phenotypes. A given application may call for emphasizing sensitivity at the expense of specificity or vice-versa. For example, in detecting BRCA1 mutations, or with aggressive diseases such as pancreatic cancer, high sensitivity is important, whereas for more common and less aggressive cancers, such as prostate, it may be preferable to limit the number of false alarms and achieve high specificity. In principle, selecting the appropriate threshold t in (1) allows one to achieve a desired tradeoff. (A disadvantage of *TSP* is the lack of a discriminant, and thus a procedure to adjust sensitivity and specificity.) It should be noted, however, that in practice estimating the threshold on the training data can be difficult; moreover, introducing a non-zero threshold makes the decision rule somewhat more difficult to interpret. As an example, Figure 2 displays the ROC curve of the *TSM* classifier for the BRCA1

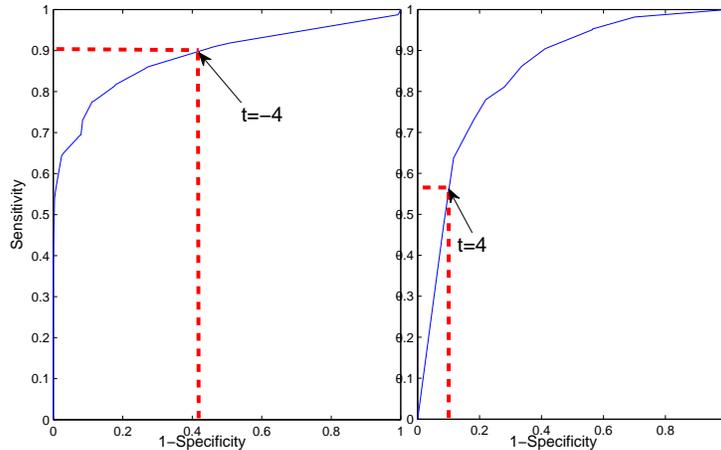


FIG 2. ROC curves for TSM. Left: BRCA1 data. With the indicated threshold, we can achieve sensitivity around 0.9 at the expense of specificity around 0.6. Right: Prostate 4 data. The given threshold reaches 0.88 specificity at the expense of sensitivity around 0.55.

and Prostate 4 studies, together with thresholds achieving hypothetically desired scenarios.

4. Experimental Results. A summary of the rank-based discriminants developed in the preceding sections is given in Table 2. We learned each discriminant for each of the datasets listed in Table 1. Among an abundance of proposed methods for high-dimensional data classification (e.g., Bradley and Mangasarian (1998), Zhang et al. (2006), Marron, Todd and Ahn (2007)), we chose two of the most effective and popular choices for predicting phenotypes from expression data: *PAM* (Tibshirani et al. (2002)), which is a form of *LDA*, and *SVM-RFE* (Guyon et al. (2002)), which is a form of linear *SVM*.

Generalization errors are estimated with cross-validation, specifically averaging the results of ten repetitions of 10-fold CV, as recommended in Braga-Neto and Dougherty (2004) and Hastie, Tibshirani and Friedman (2001). Despite the inaccuracy of small-sample cross-validation estimates (Braga-Neto and Dougherty (2004)), 10-fold CV suffices to obtain the broad perspective on relative performance across many different datasets.

The protocols for training (including parameter selection) are given below. To reduce computation, we filter the whole gene pool without using the class labels before selecting the context for rank discriminants (*TSP*, *KTSP* and *TSM*). Although a variety of filtering methods exist in the literature, such as

	Parameters	Discriminant	Parameter Selection
General	(Θ_k, k) $\Theta_k \subset \{1, \dots, d\}$	$g(X; \Theta_k)$ $\hat{\delta}(\Theta_k) = \hat{E}(g(X; \Theta_k) Y=1) - \hat{E}(g(X; \Theta_k) Y=0)$ $\hat{\sigma}(\Theta_k) = \sqrt{\widehat{\text{Var}}(g Y=0) + \widehat{\text{Var}}(g Y=1)}$	$\Theta_k^* = \arg \max_{\Theta_k} \hat{\delta}(\Theta_k)$ $k^* = \arg \max_k \frac{\hat{\delta}(\Theta_k^*)}{\hat{\sigma}(\Theta_k^*)}$
Examples			
TSP	$\Theta = (i, j)$	$g_{TSP} = I(X_i < X_j) - \frac{1}{2}$ $\hat{s}_{ij} = P(X_i < X_j Y=1) - P(X_i < X_j Y=0)$	$\Theta^* = \arg \max_{(i,j) \in \Theta} \hat{s}_{ij}$
KTSP	$\Theta_k = \{i_1, j_1, \dots, i_k, j_k\}$	$g_{KTSP} = \sum_{r=1}^k [I(X_{i_r} < X_{j_r}) - \frac{1}{2}]$	$\Theta_k^* = \arg \max_{\Theta_k} \sum_{r=1}^k \hat{s}_{i_r j_r}$
TSM	$\Theta_k = G_k^+ \cup G_k^-$ $G_k^- = \{i_1, \dots, i_k\}$ $G_k^+ = \{j_1, \dots, j_k\}$	$g_{TSM} = \text{med}_{j \in G_k^+} R_j - \text{med}_{i \in G_k^-} R_i$ R_i : rank of gene i in $G_k^+ \cup G_k^-$	$\Theta_k^* \approx \arg \max_{\Theta_k} \sum_{i \in G_k^-, j \in G_k^+} \hat{s}_{ij}$

TABLE 2

Summary of rank discriminants: *First column: the rank-based classifiers considered in this paper. Second column: the structure of the context Θ_k , the genes appearing in the classifier; For kTSP and TSM, Θ_k contains $2k$ genes. Third column: the form of the rank discriminant; the classifier is $f(X) = I(g(X) > 0)$. Fourth column: the selection of the context from training data. For a fixed k we select Θ_k to maximize $\hat{\delta}$, and then choose k to maximize $\hat{\delta}$ normalized by $\hat{\sigma}$.*

PAM (Tibshirani et al. (2002)), SIS (Fan and Lv (2008)), Dantzig selector (Candes and Tao (2007)) and the Wilcoxon-Rank test (Wilcoxon (1945)), we simply use an average signal filter: select the 4000 genes with highest mean rank (across both classes). In particular, there is no effort to detect “differentially expressed” genes. In this way we minimize the influence of the filtering method in assessing the performance of rank discriminants.

- *TSP*: The single pair which maximizing s_{ij} over all pairs in the 4000 filtered genes, breaking scoring ties if necessary with the secondary score proposed in Xu et al. (2005).
- *KTSP*: The k disjoint pairs maximizing s_{ij} over all pairs in the 4000 filtered genes with the same tie-breaking method. The number of pairs k is determined via Algorithm S1, within the range $k = 3, 5, \dots, 9$, avoiding ties in voting. Notice that $k = 1$ is excluded so that *KTSP* cannot reduce to *TSP*. We tried also $k = 3, 5, \dots, 49$ and the cross-validated accuracies changed insignificantly.
- *TSM*: The context is chosen from the top 4000 genes by the greedy selection procedure described in Algorithm S2. The size of the two sets for computing the median rank is selected in the range $k = 3, 5, 7, 9$ (providing a unique median and thereby rendering Proposition S1 applicable). We also tried $k = 3, 5, \dots, 49$ and again the changes in the

DataSet	TSP	TSM	KTSP	SVM	PAM
Colon	88/88	86/88	87/86	87/73	83/81
BRCA1	71/75	90/75	88/77	68/88	39/82
CNS	41/79	81/88	67/93	52/86	77/79
DLBCL	98/97	96/95	96/88	97/91	72/100
Lung	92/97	97/99	94/100	95/100	97/100
Marfan	82/93	89/90	88/96	99/93	88/87
Crohn's	89/90	92/91	92/96	100/100	93/98
Sarcoma	83/78	88/89	93/91	97/94	93/100
Squamous	89/88	88/85	99/92	94/95	94/95
GCM	81/73	88/77	90/75	94/80	95/94
Leukemia 1	90/85	97/94	97/93	98/97	95/89
Leukemia 2	96/96	100/93	100/96	100/96	73/88
Leukemia 3	98/98	97/99	97/98	100/100	96/99
Leukemia 4	92/94	95/98	96/97	99/97	77/92
Prostate 1	95/93	89/96	90/95	91/95	89/91
Prostate 2	68/68	76/79	76/83	68/79	77/74
Prostate 3	97/79	99/90	99/83	99/100	98/100
Prostate 4	77/61	87/70	86/79	92/62	66/85
Prostate 5	97/99	97/98	95/99	100/99	99/100
Breast 1	82/90	82/91	85/91	77/88	95/98
Breast 2	83/82	73/89	75/87	71/86	86/88

TABLE 3

Sensitivity/specificity for different classification methods. Overall accuracy is calculated as the average of sensitivity and specificity.

cross-validated accuracies were insignificant.

- *SVM-RFE*: We learned two linear *SVMs* using *SVM-RFE*: one with ten genes and one with a hundred genes. No filtering was applied, since *SVM-RFE* itself does that. Since we found that the choice of the slack variable barely changes the results, we fix $C = 0.1$. (In fact, the data are linearly separable in nearly all loops.) Only the results for *SVM-RFE* with a hundred genes are shown since it was almost 3% better than with ten genes.
- *PAM*: We use the automatic filtering mechanism provided by Tibshirani (2011). The prior class likelihoods were set to 0.5 and all other parameters were set to default values. The most important parameter is the threshold; the automatic one chosen by the program results in relatively lower accuracy than the other methods (84.00%) on average. Fixing the threshold and choosing the best one over all datasets only increases the accuracy by one percent. Instead, for each dataset and each threshold, we estimated the cross-validated accuracy for *PAM* and report the accuracy of the best threshold for that dataset.

Table 3 shows the performance estimates of the classifiers across 21 datasets.

In addition, Figures S1 and S2 in Supplement B (Afsari, Braga-Neto and Geman (2014b)) display the results in box plot format. The averages are: *TSP* (85.59%), *KTSP* (90.07%), *TSM* (88.97%), *SVM-RFE* (89.92%) and *PAM* (88.19%). The differences in the averages among methods do not appear substantial, with the possible exception of *TSP*, which lags behind the others.

There are however clearly significant variations in performance within individual datasets. In order to examine these variations at a finer scale, possibly revealing trends to support practical recommendations, recall that for each dataset and each method, we did ten repetitions of ten-fold cross-validation, resulting in one hundred trained classifiers and estimated rates (on the left-out subsets), which were averaged to provide a single cross-validated classification rate. The notch-boxes for each dataset and method are plotted in Figures S1 and S2 (Supplement B). As is commonly done, any two methods will be declared to be “tied” on a given dataset if the notches overlap; otherwise, i.e., if the notches are disjoint, the “winner” is taken to be the one with the larger median.

First, using the “notch test” to compare the three RIC classifiers, *KTSP* slightly outperforms *TSM*, which in turn outperforms *TSP*. More specifically, *KTSP* has accuracy superior to both others on ten datasets. In terms of *KTSP* vs. *TSM*, *KTSP* outperforms on three datasets, vice-versa on one dataset and they tie on all others. Moreover, *TSM* outperforms *TSP* on nine datasets and vice-versa on two datasets. As a result, if accuracy is the dominant concern, we recommend *KTSP* among the RIC classifiers, whereas if simplicity, transparency and links to biological mechanisms are important, one might prefer *TSP*. Comparisons with non-RIC methods (see below) are based on *KTSP*, although substituting *TSM* does not lead to appreciably different conclusions.

Second, *SVM* performs better than *PAM* on six datasets and *PAM* on three datasets. Hence, in the remainder of this section we will compare *KTSP* with *SVM*. We emphasize that the comparison between *PAM* and *SVM* is on our particular datasets, using our particular measures of performance, namely cross-validation to estimate accuracy and the notch test for pairwise comparisons. Results on other data sets or in other conditions may differ.

Third, whereas the overall performance statistics for *KTSP* and *SVM* are almost identical, trends do emerge based on sample size, which is obviously an important parameter and especially useful here because it varies considerably among our datasets (Table 1). To avoid fine-tuning, we only consider a coarse and somewhat arbitrary quantization into three categories: “small,” “medium” and “large” datasets, defined, respectively, by fewer than 100

(total) samples (twelve datasets), 100-200 samples (five datasets) and more than 200 samples (four datasets). On small datasets, *KTSP* outperforms *SVM* on four datasets and never vice-versa; for medium datasets each outperforms the other on one of the five datasets; and *SVM* outperforms *KTSP* on three out of four large datasets and never vice versa.

Another criterion is sparsity: the number of genes used by *TSP* is always two and by *SVM-RFE* is always one hundred. Averaged across all datasets and loops of cross-validation, *KTSP* uses 12.5 genes, *TSM* uses 10.16 genes, and *PAM* uses 5771 genes.

Finally, we performed an experiment to roughly gauge the variability in selecting the genes in the support of the various classifiers. Taking advantage of the fact that we train 100 different classifiers for each method and dataset, each time with approximately the same number of examples, we define a “consistency” measure for a pair of classifiers as the average support overlap over all distinct pairs of runs. That is, for any given dataset and method, and any two loops of cross validation, let S_1 and S_2 be the supports (set of selected genes) and define the overlap as $\frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$. This fraction is then averaged over all $100(99)/2$ pairs of loops, and obviously ranges from zero (no consistency) to one (consistency in all loops). Whereas in 16 of the 21 datasets *KTSP* had a higher consistency score than *SVM*, the more important point is that in both cases the scores are low in absolute terms, which coheres with other observations about the enormous variations in learned genes signatures.

5. Discussion and Conclusions. What might be a “mechanistic interpretation” of the *TSP* classifier, where the context consists of only two genes? In Price et al. (2007), a reversal between the two genes Prune2 and Obscurin is shown to be an accurate test for separating GIST and LMS. Providing an explanation, an hypothesized mechanism, is not straightforward, although it has been recently shown that both modulate RhoA activity (which controls many signaling events): a splice variant of Prune2 is reported to decrease RhoA activity when over-expressed and Obscurin contains a Rho-GEF binding domain which helps to activate RhoA (Funk (2012)).

Generically, one of the most elementary regulatory motifs is simply A inhibits B (denoted $A \dashv B$). For example, A may be constitutively “on” and B constitutively “off” after development. Perhaps A is a transcription factor or involved in methylation of B . In the normal phenotype we see A expressed but perhaps A becomes inactivated in the cancer phenotype, resulting in the expression of B , and hence an expression reversal from normal to cancer. Still

more generally, a variety of regulatory feedback loops have been identified in mammals. For instance, an example of a bi-stable loop is shown below.



FIG 3. A bi-stable feedback loop. Molecules A_1, A_2 (resp. B_1, B_2) are from the same species, for example two miRNAs (resp., two mRNAs). Letters in boldface indicate an “on” state.

Due to the activation and suppression patterns depicted in Figure 3, we might expect $P(X_{A_1} < X_{A_2} | Y = 0) \gg P(X_{A_1} < X_{A_2} | Y = 1)$ and $P(X_{B_1} < X_{B_2} | Y = 0) \ll P(X_{B_1} < X_{B_2} | Y = 1)$. Thus there are two expression reversals, one between the two miRNAs and one, in the opposite direction, between the two mRNAs. Given both miRNA and mRNA data, we might then build a classifier based on these two switches. For example, the rank discriminant might simply be 2TSP, the number of reversals observed. Accordingly, we have argued that expression comparisons may provide an elementary building block for a connection between rank-based decision rules and potential mechanisms.

We have reported extensive experiments with classifiers based on expression comparisons with different diseases and microarray platforms and compared the results with other methods which usually use significantly more genes. No one classifier, whether within the rank-based collection or between them and other methods such as *SVM* and *PAM*, uniformly dominates. The most appropriate one to use is likely to be problem-dependent. Moreover, until much larger datasets become available, it will be difficult to obtain highly accurate estimates of generalization errors. What does seem apparent is that our results support the conclusions reached in earlier studies (Dudoit, Fridlyand and Speed. (2002), Braga-Neto (2007), Wang (2012), Simon et al. (2003)) that simple classifiers are usually competitive with more complex ones with microarray data and limited samples. This has important consequences for future developments in functional genomics since one key thrust of “personalized medicine” is an attempt to learn appropriate treatments for disease subtypes, which means sample sizes will not necessarily get larger and might even get *smaller*. Moreover, as attention turns increasingly towards treatment, potentially mechanistic characterizations of statistical decisions will become of paramount importance for translational

medicine.

SUPPLEMENTARY MATERIAL

Proposition S1

(doi: COMPLETED BY THE TYPESETTER; .pdf). We provide the statement and proof of Proposition S1 as well as statistical tests for the assumptions made in Proposition S1.

Notch-plots for Classification Accuracies

(doi: COMPLETED BY THE TYPESETTER; .pdf). We provide notch-plots of the estimates of classification accuracy for every method and every dataset based on ten runs of ten-fold cross-validation.

Algorithms for KTSP and TSM

(doi: COMPLETED BY THE TYPESETTER; .pdf). We provide a summary of the algorithms for learning the KTSP and TSM classifiers.

References.

- AFSARI, B., BRAGA-NETO, U. and GEMAN, D. (2014a). Supplement A to: Rank Discriminants for Predicting Phenotypes from RNA Expression. *Annals of Applied Statistics*.
- AFSARI, B., BRAGA-NETO, U. and GEMAN, D. (2014b). Supplement B to: Rank Discriminants for Predicting Phenotypes from RNA Expression. *Annals of Applied Statistics*.
- AFSARI, B., BRAGA-NETO, U. and GEMAN, D. (2014c). Supplement C to: Rank Discriminants for Predicting Phenotypes from RNA Expression. *Annals of Applied Statistics*.
- ALON, U., BARKAI, N., NOTTERMAN, D. et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS* **96**.
- ALTMAN, R. B., HO K. KROEMER, C. A. M. et al. (2011). Pharmacogenomics: will the promise be fulfilled. *Nature Reviews* **12** 69-73.
- ANDERSON, T., TCHERNYSHYOV, I., DIEZ, R. et al. (2007). Discovering robust protein biomarkers for disease from relative expression reversals in 2-D DIGE data. *Proteomics* **7**.
- ARMSTRONG, S., STAUNTON, J., SILVERMAN, L. et al. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics* **30** 41-47.
- AUFFRAY, C. (2007). Protein subnetwork markers improve prediction of cancer outcome. *Molecular Systems Biology* **3**.
- BICCIATO, S., PANDIN, M., DIDON, G. and BELLO, C. D. (2003). Pattern identification and classification in gene expression data using an autoassociative neural network model. *Biotechnology Bioengineering* **81** 594-606.
- BLOATED, B., IRIZARRY, R. and SPEED, T. (2004). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19** 185-193.
- BLOOM, G., YANG, I., BOULWARE, D. et al. (2004). Multi-platform, multisite, microarray-based human tumor classification. *American Journal of Pathology* **164** 9-16.
- BOULESTEIX, A. L., TUTZ, G. and STRIMMER, K. (2003). A CART-based approach to discover emerging patterns in microarray data. *Bioinformatics* **19** 2465-2472.

- BRADLEY, P. S. and MANGASARIAN, O. L. (1998). Feature Selection via Concave Minimization and Support Vector Machines. In *ICML* 82-90.
- BRAGA-NETO, U. M. (2007). Fads and fallacies in the name of small-sample microarray classification—a highlight of misunderstanding and erroneous usage in the applications of genomic signal processing. *IEEE Signal Processing Magazine* **24** 91-99.
- BRAGA-NETO, U. M. and DOUGHERTY, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics* **20** 374-380.
- BUFFA, F., CAMPS, C., WINCHESTER, L., SNELL, C., GEE, H., SHELDON, H., TAYLOR, M., HARRIS, A. and RAGOISSIS, J. (2011). microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. *Cancer Research* **71** 5635-45.
- BURCZYNSKI, M., PETERSON, R., TWINE, N. et al. (2006). Molecular classification of Crohn’s disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *The Journal of Molecular Diagnostic* **8** 51-61.
- CANDES, E. and TAO, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n . *Annals of Statistics* **35** 2313-2351.
- CASELLA, G. and BERGER, R. L. (2002). *Statistical Inference*, 2nd ed. Duxbury, Pacific Grove, CA.
- DETLING, M. and BUHLMANN, P. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics* **19**.
- DUDOIT, S., FRIDLAND, J. and SPEED, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97** 77-87.
- EDELMAN, L., TOIA, G., GEMAN, D. et al. (2009). Two-transcript gene expression classifiers in the diagnosis and prognosis of human diseases. *BMC Genomics* **10**.
- ENERLY, E., STEINFELD, I., KLEIVI, K., LEIVONEN, S.-K. et al. (2011). miRNA-mRNA Integrated Analysis Reveals Roles for miRNAs in Primary Breast Tumors. *PLOS ONE* **6**.
- EVANS, J. P., MESLIN, E. M., MARTEAU, T. M. and CAULFIELD, T. (2011). Deflating the Genomic Bubble. *Science* **331** 861-862.
- FAN, J. and FAN, Y. (2008). High dimensional classification using features annealed independence rules. *Annals of statistics* **36** 2605.
- FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Royal Statistical Society Journal of the Royal Statistical Society Series B* **70**.
- FUNK, C. (2012). Personal Communication. Institute for Systems Biology, Seattle, WA.
- GEMAN, D., D’AVIGNON, C., NAIMAN, D. et al. (2004). Gene expression comparisons for class prediction in cancer studies. In *Proceedings 36’t Symposium on the Interface: Computing Science and Statistics*.
- GEMAN, D., AFSARI, B., TAN, A. C. and NAIMAN, D. Q. (2008). Microarray classification from several two-gene expression comparisons In *Machine Learning and Applications, 2008. ICMLA ’08. Seventh International Conference on* 583–585. IEEE (Winner, ICMLA Microarray Classification Algorithm Competition).
- GOLUB, T. R., SLONIM, D. K., TAMAYO, P. et al. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* **286** 531-537.
- GORDON, G. J., JENSEN, R. V., HSIAO, L.-L., GULLANS, S. R., BLUMENSTOCK, J. E., RAMASWAMY, S., RICHARDS, W. G., SUGARBAKER, D. J. and BUENO, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research* **62** 4963-4967.
- GUO, Y., HASTIE, T. and TIBSHIRANI, R. (2005). Regularized discriminant analysis and

- its application in microarrays. *Biostatistics* **1** 1–18.
- GUO, Y., HASTIE, T. and TIBSHIRANI, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8** 86–100.
- GUYON, I., WESTON, J., BARNHILL, S. and VAPNIK, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **46** 389–422.
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- JONES, S., ZHANG, X., PARSONS, D. W. et al. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321** 1801–1806.
- KHAN, J., WEI, J. S., RINGNR, M. et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* **7**.
- KOHLMANN, A., KIPPS, T. J., RASSENTI, L. Z., DOWNING, J. R., , et al. (2008). An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the Microarray Innovations in LEukemia study prephase. *British journal of haematology* **142** 802–807.
- KURIAKOSE, M. A., CHEN, W. T. et al. (2004). Selection and validation of differentially expressed genes in head and neck cancer. *Cellular and Molecular Life Sciences* **61** 13721383.
- LEE, E., CHUANG, H. Y., KIM, J. W. et al. (2008). Inferring Pathway Activity toward Precise Disease Classification. *PLOS Computational Biology* **4**.
- LIN, X., AFSARI, B., MARCHIONNI, L. et al. (2009). The ordering of expression among a few genes can provide simple cancer biomarkers and signal BRCA1 mutations. *BMC Bioinformatics* **10**.
- MA, X. J., WANG, Z., RYAN, P. D. et al. (2004). A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* **5**.
- MARRON, J., TODD, M. J. and AHN, J. (2007). Distance-weighted discrimination. *Journal of the American Statistical Association* **102** 1267–1271.
- MARSHALL, E. (2011). Waiting for the Revolution. *Science* **331** 526–529.
- MILLS, K., KOHLMANN, A., WILLIAMS, P., WIECZOREK, L. et al. (2009). Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of AML transformation of myelodysplastic syndrome. *Blood*.
- PENG, S., XU, Q., LING, X. et al. (2003). Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS Letters* **555** 358–362.
- POMEROY, C., TAMAYO, P., GAASENBEEK, M. et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415** 436–442.
- PRICE, N., TRENT, J., EL-NAGGAR, A. et al. (2007). Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas. *PNAS* **43**.
- QU, Y., ADAM, B., YASUI, Y. et al. (2002). Boosted decision tree analysis of surface-enhanced laser desorption/ ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clinical Chemistry* **48** 1835–1843.
- RAMASWAMY, S., TAMAYO, P., RIFKIN, R. et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *PNAS* **98** 15149–15154.
- RAPONI, M., LANCET, J. E., FAN, H. et al. (2008). A 2-gene classifier for predicting response to the farnesyltransferase inhibitor tipifarnib in acute myeloid leukemia. *Blood* **111** 2589–2596.
- SHIPP, M., ROSS, K., TAMAYO, P. et al. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* **8** 68–74.

- SIMON, R., RADMACHER, M. D., DOBBIN, K. and MCSHANE, L. M. (2003). Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. *Journal of the National Cancer Institute* **95** 14-18.
- SINGH, D., FEBBO, P., ROSS, K. et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1** 203-209.
- STUART, R., WACHSMAN, W., BERRY, C. et al. (2004). In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. *PNAS* **101** 615-620.
- TAN, A. C., NAIMAN, D. Q., XU, L. et al. (2005). Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* **21** 3896-3904.
- THOMAS, R. K., BAKER, A. C., DEBIASI, R. M. et al. (2007). High-throughput oncogene mutation profiling in human cancer. *Nature Genetics* **39** 347-351.
- TIBSHIRANI, R. (2011). PAM R Package. <http://www-stat.stanford.edu/~tibs/PAM/Rdist/index.html>.
- TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. and CHU, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS* **99** 6567-6572.
- WANG, X. (2012). Robust two-gene classifiers for cancer prediction. *Genomics* **99** 90-95.
- WEICHELBAUM, R. R., ISHWARANC, H., YOONA, T. et al. (2008). An interferon-related gene signature for DNA damage resistance is a predictive marker for chemotherapy and radiation for breast cancer. *PNAS* **105** 1849018495.
- WELSH, J., SAPINOSO, L., SU, A. et al. (2001). Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res* **61** 5974-5978.
- WILCOXON, F. (1945). Individual comparisons by ranking methods. *Biometrics* 80-83.
- WINSLOW, R., TRAYANOVA, N., GEMAN, D. and MILLER, M. (2012). The emerging discipline of computational medicine. *Science Translational Medicine* **4** 158rv11.
- XU, L., GEMAN, D. and WINSLOW, R. L. (2007). Large-scale integration of cancer microarray data identifies a robust common cancer signature. *BMC Bioinformatics* **8**.
- XU, L., TAN, A. C., NAIMAN, D. Q. et al. (2005). Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *BMC Bioinformatics* **21** 39053911.
- YAO, Z., JAEGER, J., RUZZO, W. L. et al. (2004). Gene Expression Alterations in Prostate Cancer Predicting Tumor Aggression and Preceding Development of Malignancy. *Journal of Clinical Oncology* **22** 2790-2799.
- YAO, Z., JAEGER, J., RUZZO, W. et al. (2007). A Marfan syndrome gene expression phenotype in cultured skin fibroblasts. *BMC Genomics* **8**.
- YEANG, C., RAMASWAMY, S., TAMAYO, P. et al. (2001). Molecular classification of multiple tumor types. *Bioinformatics* **17**.
- ZHANG, H., YU, C. Y. and SINGER, B. (2003). Cell and tumor classification using. *PNAS* **100** 4168-4172.
- ZHANG, H. H., AHN, J., LIN, X. and PARK, C. (2006). Gene selection using support vector machines with non-convex penalty. *Bioinformatics* **22**.
- ZHAO, H., LOGOTHETIS, C. J. and GORLOV, I. P. (2010). Usefulness of the top-scoring pairs of genes for prediction of prostate cancer progression. *Prostate Cancer and Prostatic Diseases* **13** 252-259.

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
WHITING SCHOOL OF ENGINEERING
JOHNS HOPKINS UNIVERSITY
bahman@jhu.edu

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING AND
CENTER FOR BIOINFORMATICS AND GENOMICS SYSTEMS ENGINEERING
DWIGHT LOOK COLLEGE OF ENGINEERING
TEXAS A&M UNIVERSITY
ub@ieee.org

DEPARTMENT OF APPLIED MATHEMATICS AND STATISTICS
WHITING SCHOOL OF ENGINEERING
JOHNS HOPKINS UNIVERSITY
geman@jhu.com