JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 11, Number 6, 2004

Pp. 1073–1089

Robust Sparse Hyperplane Classifiers: Application to Uncertain Molecular Profiling Data

C. BHATTACHARYYA,[1,4,5] L.R. GRATE,[2,5] M.I. JORDAN,[1,3] L. EL GHAOUI,[1] and I.S. MIAN[2]

ABSTRACT

Molecular profiling studies can generate abundance measurements for thousands of transcripts, proteins, metabolites, or other species in, for example, normal and tumor tissue samples. Treating such measurements as features and the samples as labeled data points, sparse hyperplanes provide a statistical methodology for classifying data points into one of two categories (classification and prediction) *and* defining a small subset of discriminatory features (relevant feature identification). However, this and other extant classification methods address only implicitly the issue of observed data being a combination of underlying signals and noise. Recently, robust optimization has emerged as a powerful framework for handling uncertain data explicitly. Here, ideas from this field are exploited to develop *robust* sparse hyperplanes, i.e., classification and relevant feature identification algorithms that are resilient to variation in the data. Specifically, each data point is associated with an explicit data uncertainty model in the form of an ellipsoid parameterized by a center and covariance matrix. The task of learning a robust sparse hyperplane from such data is formulated as a second order cone program (SOCP). Gaussian and distribution-free data uncertainty models are shown to yield SOCPs that are equivalent to the SCOP based on ellipsoidal uncertainty. The real-world utility of robust sparse hyperplanes is demonstrated via retrospective analysis of breast cancer related transcript profiles. Data-dependent heuristics are used to compute the parameters of each ellipsoidal data uncertainty model. The generalization performance of a specific implementation, designated "robust LIKNON," is better than its nominal counterpart. Finally, the strengths and limitations of robust sparse hyperplanes are discussed.

Key words: robust sparse hyperplanes, second-order cone program, linear programming, breast cancer, molecular profiling, two-class high-dimensional data.

[1]Department of EECS, University of California Berkeley, Berkeley, CA 94720.
[2]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720.
[3]Department of Statistics, University of California Berkeley, Berkeley, CA 94720.
[4]*Current address*: Department of CSA, Indian Institute of Science, Bangalore 560012, Karnataka, India.
[5]These authors contributed equally.

1. INTRODUCTION

TRANSCRIPT, PROTEIN, METABOLITE, or other molecular profiling technologies yield data that pose challenges for biologists and statisticians alike. In transcriptional profiling, currently the most mature and widely deployed technology (Golub *et al.*, 1999; Bhattacharjee *et al.*, 2001; Garber *et al.*, 2001; Sørlie *et al.*, 2001; Hedenfalk *et al.*, 2001; Notterman *et al.*, 2001; Dhanasekaran *et al.*, 2001; Ramaswamy *et al.*, 2001; Su *et al.*, 2001; Khan *et al.*, 2001; Liotta *et al.*, 2001), the abundances of nucleic acid sequences (genes for brevity) are monitored in biological samples of interest. Frequently, the number of molecular species measured exceeds the number of samples assayed by one to two orders of magnitude. Thus, the input for subsequent analysis is high-dimensional data where each feature (gene, protein, metabolite, and so on) of a data point (sample) represents the abundance of a species.

This work considers scenarios in which data points are assigned to one of two categories. Given two-class high-dimensional data, one biological question is ascertaining a small number of features able to distinguish different classes of data points. *Classification* and *prediction* methods estimate a model from data and employ the learned system to assign the class of a previously unseen data point. For example, a classifier trained using breast tissue samples could presage a clinical decision support system designed to predict whether a new patient sample was or was not normal. *Relevant feature identification* methods define a subset of the features able to discriminate between classes. Enumerating genes able to distinguish normal from malignant samples could portend novel and/or improved targets for intervention, diagnosis, and imaging (biomarkers).

Sparse classifiers are statistical algorithms for classification (differentiating two classes of data points) and relevant feature identification (specifying a small subset of discriminatory features). For linearly separable data, the decision surface separating the classes is a hyperplane parameterized by a weight vector and an offset term. A sparse weight vector, one with few nonzero elements, specifies the relevant features. Sparse hyperplanes can be estimated from data by formulating an optimization problem as a linear program (LP) that minimizes an l_1 norm (Donoho and Huo, 1999; Graepel *et al.*, 1999; Smola *et al.*, 1999; Bennett and Demiriz, 1999; Cristianini and Shawe-Taylor, 2000; Hastie *et al.*, 2000; Bennett and Campbell, 2000).

Nominal LIKNON is a specific implementation of nominal sparse hyperplanes that has been applied to a variety of real-world transcript and protein profiles (Bhattacharyya *et al.*, 2003; Grate *et al.*, 2003, 2002). Computationally, it had nontrivial advantages over the prevailing multistep filter-wrapper strategy because one pass through data yielded both a classifier and relevant features. Biologically, the distinguishing features suggested gene biomarkers for a number of cancers. Subsequent hidden Markov model-based sequence analysis of the protein product of one of these biomarkers resulted in the definition and characterization of a novel, phylogenetically conserved protein family (Grate *et al.*, 2003).

Nonrandom and random processes give rise to uncertain profiling data. Observed abundance measurements can be viewed as signals that are a convolution of the (patho-)biology or stimulus under investigation, and technical factors (see, for example, Novak *et al.* [2002]). However, extant classification methods such as nominal sparse hyperplanes account for biological and experimental variability only implicitly. Thus, nominal LIKNON yields a linear classifier and relevant features that may not be immune to variation in the data. This work proposes robust sparse hyperplanes as classification and relevant feature identification models that are resilient to this key facet of real-world data. Realizing these models requires building upon and exploiting recent advances in the field of robust optimization (Ben-Tal *et al.*, 2000; Boyd and Vandenberghe, 2003), an area underexplored in computational biology.

In our approach, the notion of "uncertainty" is made explicit by specifying the allowable values of a data point via an ellipsoidal data uncertainty model parameterized by a center (location) and covariance matrix (shape). The task of learning a robust sparse hyperplane from such models is posed as a robust LP (Ben-Tal *et al.*, 2000) and a second order cone program (SOCP). SOCPs are a class of nonlinear convex optimization problems that can be solved efficiently using interior-point algorithms and methods and, like LPs, have single global solutions (see Boyd and Vandenberghe [2003]). Two other data uncertainty models, Gaussian or distribution-free, yield equivalent SOCPs. Robust LIKNON is a specific implementation of robust sparse hyperplanes based on ellipsoidal data uncertainty models. The size of the SOCP to be solved is reduced by assuming that some data points are associated with identically shaped uncertainty: one covariance matrix per data point is replaced by one (class-independent) or two (class-dependent) matrices per dataset.

The utility of robust sparse hyperplanes is illustrated using published transcript profiles for 3,226 genes in 7 BRCA1 and 15 BRCA2/sporadic breast tumor samples (Hedenfalk *et al.*, 2001). Since repeated

measurements were not available, data-dependent heuristics were employed to specify the 22 ellipsoid centers and to compute a feature-dependent diagonal covariance matrix for points in the dataset (22) or each category (7, 15). To investigate the effect of different amounts of data uncertainty, the covariance matrix was scaled via a user-defined noise level parameter ρ. Robust LIKNON hyperplanes estimated across a range of noise values all specified a small number of relevant features (10–20) and yielded good linear classifiers (low or zero classification error). Hyperplanes in which the shape of the uncertainty associated with each category was unique (two class-dependent covariance matrices) were more robust than ones where all points had identically shaped uncertainty (one class-independent matrix). On a standard computer, typical runtimes were a few minutes and the memory requirements were small.

This paper concludes with a discussion of implementation and model selection (number of relevant features) issues. In addition, robust sparse hyperplanes are compared with a recent heuristic Monte Carlo–based strategy for estimating a linear classifier and defining a small number of informative features from uncertain data (Kim *et al.*, 2002).

2. MATERIALS AND METHODS

2.1. Transcriptional profiling data

Published cDNA microarray data for BRCA1 mutation positive, BRCA2 mutation positive and sporadic breast tumor samples (Hedenfalk *et al.*, 2001) were downloaded from *www.nhgri.nih.gov/DIR/Microarray/NEJM_Supplement.* The transcript profile for a sample consists of log_2 transformed ratios of (background corrected) fluorescent intensities for genes in the sample of interest versus a reference sample. There were 3,226 genes (features) and 22 samples (data points). The 7 BRCA1 samples were designated the $+1$ class and the 15 BRCA2 and sporadic samples the -1 class.

2.2. Sparse hyperplanes: Nominal LIKNON

Consider a two-class dataset consisting of N data points, $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$. Each data point is a P-dimensional vector of features, $\mathbf{x}_n \in \mathbb{R}^P$, assigned to one of two categories, $y_n \in \{+1, -1\}$. Assume that the classes are linearly separable. A classifier for such data is a hyperplane parameterized by a weight vector, $\mathbf{w} \in \mathbb{R}^P$, and an offset from the origin, $b \in \mathbb{R}$. The hyperplane $\mathcal{H}(\mathbf{w}, b)$ can be used to predict the class of a data point $\mathbf{x} \in \mathbb{R}^P$ by computing $\text{sign}(\mathbf{w}^T\mathbf{x}+b)$. If this value is positive, $\mathbf{x}$ is identified with the $+1$ class; otherwise, it belongs to the -1 class. Points that define the hyperplane, positive half-space ($+1$ class), and negative half-space (-1 class) are the sets $\{\mathbf{x}|\mathbf{w}^T\mathbf{x} = b\}$, $\{\mathbf{x}|\mathbf{w}^T\mathbf{x} > b\}$, and $\{\mathbf{x}|\mathbf{w}^T\mathbf{x} < b\}$, respectively.

A sparse classifier addresses the dual problem of discriminating between classes and identifying a small number of relevant features. For a sparse hyperplane, this translates to a weight vector $\mathbf{w}$ with few nonzero elements. The rationale is that if the pth element of $\mathbf{w}$ is nonzero, $w^p \neq 0$, the corresponding feature x^p contributes to $\text{sign}(\mathbf{w}^T\mathbf{x}+b)$ and hence the class of point $\mathbf{x}$. When $w^p = 0$, x^p plays no role in determining the class. Although many single features could differentiate the classes on their own, the ensuing classifiers are unlikely to generalize well to previously unseen examples (overfitting). Thus, the goal is to find a sparse weight vector with a small, as opposed to the smallest, number of nonzero elements.

2.2.1. Norm minimization. The problem of estimating a sparse hyperplane from data can be addressed by minimizing the l_0 norm of the weight vector $\mathbf{w}$ subject to a requirement that the classification error be low. The l_0 norm of a vector is the number of nonzero elements, $\|\mathbf{w}\|_0 = \text{cardinality}\{p : w^p \neq 0\}$. However, minimizing the l_0 norm of high-dimensional vectors is NP-hard (Amaldi and Kann, 1998). A computationally tractable approximation to minimizing the l_0 norm of a vector is to minimize its l_1 norm (Donoho and Huo, 1999) (see Weston *et al.* [2003] for a discussion of zero-norms and linear models). The l_1 norm is the sum of the absolute magnitudes of its elements, $\|\mathbf{w}\|_1 = \sum_{p=1}^P |w^p|$.

The basic optimization problem for learning a sparse hyperplane from training data is then

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \|\mathbf{w}\|_1 + C\sum_{n=1}^N \xi_n \\ \text{subject to} \quad & y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq \gamma - \xi_n \\ & \gamma = 1, \quad \xi_n \geq 0, \quad n = 1, \ldots, N. \end{aligned} \tag{1}$$

The slack variables, ξ_n, provide a mechanism for handling an error in the assigned class, y_n. This quantity measures how much the point $\mathbf{x}_n$ fails to have a margin of γ from the hyperplane, $\xi_n = \max\{0, \gamma - y_n(\mathbf{w}^T\mathbf{x}_n + b)\}$. When $\xi_n > \gamma$, the point is misclassified. The number of mistakes made by a hyperplane on training data is the number of data points for which the slack variable exceeds the margin. Since all values scale with the margin, it must be greater than zero to avoid the minimal, but not useful, all-zero solution $|\mathbf{w}|_1 = 0$. Typically, in this sort of optimization work, γ is set to 1, as in (1).

The user-defined regularization parameter, C, weighs the contribution of misclassifications (if any). Larger values result in fewer errors being ignored and a desire for good training error. Parameter $C \to 0$ indicates little penalty for misclassification whereas $C \to \infty$ is equivalent to a hard margin limit where no data point is ignored. Together, ξ_n and C influence the sparseness of the solution.

2.2.2. Nominal LIKNON. Nominal LIKNON is a specific implementation of nominal sparse hyperplanes that has been discussed elsewhere (Bhattacharyya *et al.*, 2003; Grate *et al.*, 2003, 2002). Briefly, the attractive aspects of optimization problem (1) are (i) it can be cast as an LP so there are no local minima and an optimum solution can be found and (ii) efficient algorithms for solving LPs involving ~10,000 variables and ~10,000 constraints are available. Versions of nominal LIKNON based upon on the stand-alone open source code lpsolve (*www.netlib.org/ampl/solvers/lpsolve/*) or Matlab (*www.themathworks.com*) are available at *www.cs.berkeley.edu/ jordan/liknon/*.

The value of the margin influences the magnitude of the hyperplane with larger (smaller) γ values resulting in larger (smaller) weight vectors $\mathbf{w}$. In practice, limitations in the numerical precision of solvers may lead to "hidden implementation errors." To ameliorate such potential problems, an important preprocessing step is to scale the data into a reasonable range.

2.3. Robust sparse hyperplanes: Robust LIKNON

A robust sparse classifier is one in which the decision boundary and set of relevant features are resilient to uncertainty in the data. As a consequence of the margin $\gamma = 1$, a solution to optimization problem (1) does yield a sparse hyperplane with some intrinsic robustness. However, the robust optimization paradigm provides a formal, explicit framework for handling data uncertainty and thus a principled approach to estimating robust sparse hyperplanes.

2.3.1. Data uncertainty model and robust LPs. Assume that the observed and all allowable values of a data point can be described by a data uncertainty model; an uncertainty set $\mathcal{U}$ is the collection of values specified by this model. The robust counterpart of the sparse hyperplane optimization problem (1) stipulates that all realizations of the data, every admissible value $\mathbf{x}_{in}$ in each uncertainty set $\mathcal{U}_n$, satisfy the inequality constraint

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \|\mathbf{w}\|_1 + C\sum_{n=1}^{N}\xi_n \\ \text{subject to} \quad & y_n(\mathbf{w}^T\mathbf{x}_{in} + b) \geq 1 - \xi_n \\ \text{for all} \quad & \mathbf{x}_{in} \in \mathcal{U}_n \\ & \xi_n \geq 0, \quad n = 1, \ldots, N. \end{aligned} \tag{2}$$

The robust LP (2) could be solved by instantiating admissible values that approximate every uncertainty set and incorporating each ensuing "pseudopoint" as an additional, explicit linear constraint. To estimate a sparse hyperplane that was robust and generalized well, each $\mathcal{U}_n$ would need to be specified by a large corpus of pseudopoints. Results (data not shown) indicate that adding a few additional linear constraints to the sparse hyperplane formulation does improve performance. However, an optimization problem with far more constraints would be impractical because the resulting LP would be too large.

2.3.2. Ellipsoidal data uncertainty model and SOCPs. The task of estimating a robust sparse hyperplane can be simplified by enunciating a data uncertainty model and converting multiple linear constraints into a single nonlinear constraint. This work considers a simple ellipsoidal data uncertainty model and recasts the robust LP (2) as an SOCP. Conceptually, each feature (gene) is allowed a "noise range" around

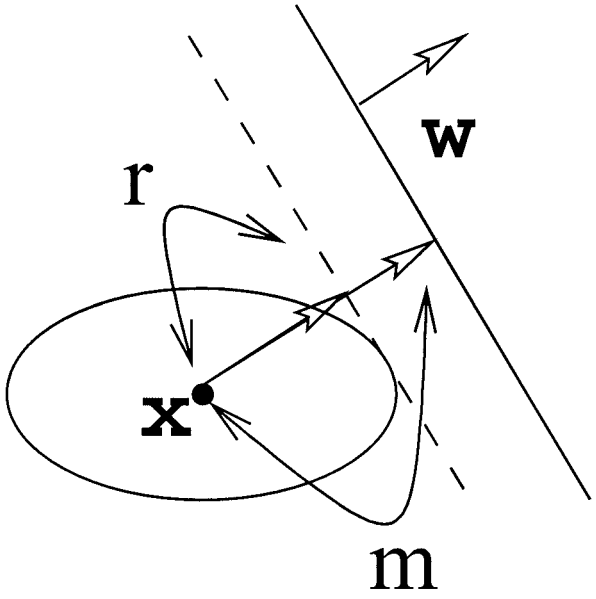

FIG. 1. The intersection of an ellipsoid, $\mathcal{E}(\tilde{\mathbf{x}}, \Sigma)$, and a hyperplane, $\mathcal{H}(\mathbf{w}, b)$. The dotted line is parallel to the hyperplane and is a tangent to the ellipse. The normal distance from $\mathbf{x}$ to the hyperplane is $m = |\mathbf{w}\mathbf{x} + b|/\|\mathbf{w}\|_2$. The distance from $\mathbf{x}$ to the dotted line is $r = (\mathbf{w}^T \Sigma \mathbf{w})^{1/2}/\|\mathbf{w}\|_2$. When $m = r$, the hyperplane is tangent to the ellipse and when $m \leq r$ the hyperplane intersects the ellipse.

the measured value, which leads to the geometrical concept of a high-dimensional ellipsoid centered on the point specified by the vector of measured values. Specifically, the uncertainty set $\mathcal{U}$ is defined by a bounded ellipsoid parameterized by two second-order statistics, a location, $\tilde{\mathbf{x}} \in \mathbb{R}^P$, and a shape, $\mathcal{R} \in \mathbb{R}^{P \times P}$,

$$\begin{aligned} \mathcal{U} &:= \{\mathbf{x}_i | (\mathbf{x}_i - \tilde{\mathbf{x}}) \mathcal{R}^2 (\mathbf{x}_i - \tilde{\mathbf{x}}) \leq 1\} \\ &\equiv \{\mathbf{x}_i | \|\mathcal{R}(\mathbf{x}_i - \tilde{\mathbf{x}})\|_2 \leq 1\}, \end{aligned} \tag{3}$$

where $\mathbf{x}_i$ is an admissible value. The l_2 norm of a vector $\mathbf{w}$ is $\|\mathbf{w}\|_2 = (\mathbf{w}^T \mathbf{w})^{1/2}$. The shape can be interpreted as a covariance matrix, Σ, that is the matrix of squared axis lengths defining the ellipsoid, $\mathcal{R} = \sqrt{\Sigma}$; the location, $\tilde{\mathbf{x}}$, can be equated with the expected value (center).

An optimal solution to the robust sparse hyperplane optimization problem (2) is one in which the hyperplane, $\mathcal{H}(\mathbf{w}, b)$, does not intersect any ellipsoidal data uncertainty model, $\mathcal{E}(\tilde{\mathbf{x}}, \Sigma)$. This is true if the following inequality holds for every ellipsoid:

$$|\mathbf{w}^T \tilde{\mathbf{x}} + b| \geq \|\Sigma^{1/2} \mathbf{w}\|_2. \tag{4}$$

The left hand side is the distance between the ellipsoid center $\tilde{\mathbf{x}}$ and the hyperplane (m in Fig. 1); the right hand side is the distance between $\tilde{\mathbf{x}}$ and a line parallel to the hyperplane and tangential to the ellipsoid (r in Fig. 1).

Inequality (4) can be used to collapse separate constraints for each admissible value in the robust LP (2) into one term involving the l_2 norm of the weight vector,

$$\begin{aligned} \min_{\mathbf{w}, \xi} \quad & \|\mathbf{w}\|_1 + C \textstyle\sum_{n=1}^{N} \xi_n \\ \text{subject to} \quad & y_n(\mathbf{w}^T \tilde{\mathbf{x}}_n + b) \geq 1 - \xi_n \\ & y_n(\mathbf{w}^T \tilde{\mathbf{x}}_n + b) \geq \|\Sigma_n^{1/2} \mathbf{w}\|_2 - \xi_n \\ & \xi_n \geq 0 \quad n = 1, \ldots, N. \end{aligned} \tag{5}$$

The nonlinear l_2 constraint prevails when the ellipsoidal uncertainty exceeds the margin, $\|\Sigma_n^{1/2} \mathbf{w}\|_2 > 1$.

Optimization problem (5) is not linear and so is no longer an LP. However, it is an instance of an SOCP. This class of convex optimization problems has the general form

$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbf{e}^T\mathbf{x} \\ \text{subject to} \quad & \mathbf{c}_i^T\mathbf{x} + d_i \geq \|A_i\mathbf{x} + b_i\|_2 \\ & i = 1, \ldots, I \end{aligned}$$

where A_i is a matrix, $\mathbf{c}_i$ and $\mathbf{x}$ are vectors, and b_i and d_i are scalars. Efficient interior-point algorithms for solving SOCPs involving thousands of variables and constraints are available.

A convenient way to rewrite (5) is to introduce an auxiliary variable t that bounds the nonlinear constraint,

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \|\mathbf{w}\|_1 + C\sum_{n=1}^{N}\xi_n \\ \text{subject to} \quad & y_n(\mathbf{w}^T\tilde{\mathbf{x}}_n + b) \geq 1 - \xi_n \\ & y_n(\mathbf{w}^T\tilde{\mathbf{x}}_n + b) \geq t_n - \xi_n \\ & \|\Sigma_n^{1/2}\mathbf{w}\|_2 \leq t_n \\ & t_n \geq 0 \\ & \xi_n \geq 0 \quad n = 1, \ldots, N. \end{aligned} \tag{6}$$

2.3.3. Robust LIKNON. A robust sparse hyperplane can be estimated from data $\{\mathcal{E}_n(\tilde{\mathbf{x}}_n, \Sigma_n), y_n\}_{n=1}^{N}$ using the SOCP (6). Since every one of the N data points has its own covariance matrix Σ_n, this formulation results in a large optimization problem (needing $NP \times P$ matrices or, if Σ_n is diagonal, $N * P$ values). Such problems are computationally too costly for current SOCP solvers (which have limitations that are now being addressed) so at present it is necessary to impose some restrictions on the shape matrices. The size of the SOCP to be solved can be reduced considerably by assuming that the shape of the uncertainty for a defined set of points is identical.

The use of class-dependent covariance matrices results in an SOCP with two data uncertainty constraints. When data points in the same category are assumed to share a common matrix, i.e., $\Sigma_n = \Sigma_+$ if $y_n = +1$ or $\Sigma_n = \Sigma_-$ if $y_n = -1$, the optimization problem becomes

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \|\mathbf{w}\|_1 + C\sum_{n=1}^{N}\xi_n \\ \text{subject to} \quad & y_n(\mathbf{w}^T\tilde{\mathbf{x}}_n + b) \geq 1 - \xi_n \\ & y_n(\mathbf{w}^T\tilde{\mathbf{x}}_n + b) \geq t_+ - \xi_n, \quad y_n = +1 \\ & y_n(\mathbf{w}^T\tilde{\mathbf{x}}_n + b) \geq t_- - \xi_n, \quad y_n = -1 \\ & \|\Sigma_+^{1/2}\mathbf{w}\|_2 \leq t_+ \\ & \|\Sigma_-^{1/2}\mathbf{w}\|_2 \leq t_- \\ & t_+ \geq 0, \quad t_- \geq 0 \\ & \xi_n \geq 0 \quad n = 1, \ldots, N. \end{aligned} \tag{7}$$

A class-independent covariance matrix results in an SOCP with one data uncertainty constraint. When all data points are presumed to share the same matrix, $\Sigma_\pm$ for $y_n = \{+1, -1\}$, the optimization problem becomes

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \|\mathbf{w}\|_1 + C\sum_{n=1}^{N}\xi_n \\ \text{subject to} \quad & y_n(\mathbf{w}^T\tilde{\mathbf{x}}_n + b) \geq 1 - \xi_n \\ & y_n(\mathbf{w}^T\tilde{\mathbf{x}}_n + b) \geq t - \xi_n \\ & \|\Sigma_{\pm}^{1/2}\mathbf{w}\|_2 \leq t \\ & t \geq 0 \\ & \xi_n \geq 0, \quad n = 1, \ldots, N. \end{aligned} \tag{8}$$

Problems (7) and (8) contain almost identical linear constraints for each data point. Combining them yields formulations with reduced memory requirements,

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \|\mathbf{w}\|_1 + C\sum_{n=1}^{N}\xi_n \\ \text{subject to} \quad & y_n(\mathbf{w}^T\tilde{\mathbf{x}}_n + b) \geq t_+ - \xi_n, \quad y_n = +1 \\ & y_n(\mathbf{w}^T\tilde{\mathbf{x}}_n + b) \geq t_- - \xi_n, \quad y_n = -1 \\ & \|\Sigma_{+}^{1/2}\mathbf{w}\|_2 \leq t_+ \\ & \|\Sigma_{-}^{1/2}\mathbf{w}\|_2 \leq t_- \\ & t_+ \geq 1, \quad t_- \geq 1 \\ & \xi_n \geq 0 \quad n = 1, \ldots, N, \end{aligned} \tag{9}$$

and

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \|\mathbf{w}\|_1 + C\sum_{n=1}^{N}\xi_n \\ \text{subject to} \quad & y_n(\mathbf{w}^T\tilde{\mathbf{x}}_n + b) \geq t - \xi_n \\ & \|\Sigma_{\pm}^{1/2}\mathbf{w}\|_2 \leq t \\ & t \geq 1 \\ & \xi_n \geq 0 \quad n = 1, \ldots, N. \end{aligned} \tag{10}$$

Robust LIKNON is a specific implementation of the above SOCPs. The software employs the extant, stand-alone, open source SOCP solver SeDuMi (Sturm, 1999), and Matlab (*http://www.mathworks.com*); the code is available as supplementary material. These robust sparse hyperplane problems revert to the nominal sparse hyperplane problem (1) when the data have no associated uncertainty, i.e., the nonlinear ellipsoid terms vanish when $\|\Sigma\|_2 = 0$.

The appendix provides a statistical interpretation of robust LPs when the data are assumed to be Gaussian or distribution-free random variables. In both cases, the ensuing optimization constraints are the same as those based on ellipsoidal data uncertainty models.

2.4. Ellipsoidal data uncertainty model parameters: Location $\tilde{\mathbf{x}}$ and shape Σ

Estimating the parameters of a robust sparse hyperplane from data is distinct and separate from ascertaining the parameters of each ellipsoidal data uncertainty model. Currently, robust LIKNON solves SOCPs based on class-dependent (Σ_+, Σ_-, (7)) and class-independent ($\Sigma_\pm$, (8)) covariance matrices. Here, each matrix is assumed to be diagonal so only the $\Sigma^{11}, \ldots, \Sigma^{PP}$ terms need to be specified a priori. In a feature-dependent diagonal covariance matrix, the Σ^{pp} terms are unique. In a feature-independent matrix, all Σ^{pp} terms are identical (this converts ellipsoids into same-radius spheres).

2.4.1. Datasets without replicates. Given data in which a biological sample is assayed only once, the prevailing scenario in current molecular profiling studies, robust LIKNON implements the following data-dependent heuristics to set the location and shape parameter values for each ellipsoidal data uncertainty model.

The ellipsoid center is equated with an observed data point, $\tilde{\mathbf{x}}_j \equiv \mathbf{x}_j$.

A covariance matrix is computed using the observed range of features in J data points. If x_j^p is the measured value of feature p in data point j, its variation, a^p, is

$$a^p = \{\max_j x_j^p - \min_j x_j^p\}, \quad j = 1, \ldots, J. \tag{11}$$

For Σ_+, Σ_-, and $\Sigma_\pm$, J is the number of data points in the +1 class, -1 class, and dataset, respectively.

In a feature-dependent (diagonal) covariance matrix, the shape of the uncertainty for feature p is set using its variation,

$$\Sigma^{pp} = a^p, \tag{12}$$

and its standard deviation is

$$\Sigma^p = \frac{1}{J}\sum_{j=1}^{J}(x_j^p)^2 - \left(\frac{1}{J}\sum_{j=1}^{J}x_j^p\right)^2. \tag{13}$$

The shape of the uncertainty can be set directly,

$$\Sigma^{pp} = \Sigma^p, \tag{14}$$

or in a globally scaled manner,

$$\Sigma^{pp} = \frac{\sum_{p=1}^{P}|a^p|}{\sqrt{\sum_{p=1}^{P}(\Sigma^p)(\Sigma^p)}}\Sigma^p. \tag{15}$$

In a feature-independent (diagonal) matrix, the shape of the uncertainty for all features is set to that for the feature having the smallest variation,

$$\Sigma^{pp} = \sqrt{P}\{\min_p a^p\}, \quad p = 1, \ldots, P. \tag{16}$$

2.4.2. Datasets with replicates. If the same biological sample is assayed multiple times, the parameters can be computed from the empirical distribution of the data. Given R independent measurements of feature p in the same sample, the location and shape can be set using

$$\tilde{x}^p = \frac{\sum_{r=1}^{R}x_r^p}{R}$$

$$\Sigma^{pp} = \frac{1}{R}\Sigma_{r=1}^{R}(x_r^p)^2 - \left(\frac{1}{R}\Sigma_{r=1}^{R}x_r^p\right)^2. \tag{17}$$

2.5. User-defined noise level parameter ρ

Computational experiments designed to evaluate the performance of robust LIKNON require datasets in which the level of variability associated with the data can be quantified. Here, a noise level parameter, $0 \leq \rho \leq 1$, is introduced to scale each diagonal element of the covariance matrix, $\rho\Sigma^{pp}$. When $\rho = 0$, data points are associated with no noise (the nominal LIKNON case). The ρ value acts as a proxy for data variability.

2.6. Sparse hyperplane regularization parameter C

An attractive aspect of the sparse hyperplane formulations is the presence of only one free parameter, the regularization parameter C. In these optimization problems, C and the slack variables ξ perform the dual function of handling nonlinearly separable data and tuning the sparsity of the solution. Both appear in the objective function, but only ξ appears in the constraints.

To obtain a feasible solution to an optimization problem, all the constraints need to be satisfied. In practice, the precise C value, $0.0 \leq C \leq \infty$, plays a critical role in determining whether this is true and in determining the sparsity of the final weight vector $\mathbf{w}$. Recall that if the slack variable is greater than the margin, $\xi_n > 1$, the point $\mathbf{x}_n$ lies on the "wrong" side of the hyperplane and is misclassified. If $0 \leq \xi_n \leq 1$, the point is classified correctly.

When the data are not linearly separable, C should be small to allow ξ, to exceed the margin if necessary. When $C = 0$, there is no control over ξ, and no solution is found. If $C = \infty$, the slack variables are forced to zero, and no misclassifications are allowed. A good solution lies between these lower and upper bounds. A hyperplane that makes few errors and has a small number of nonzero elements can be found by increasing C above zero. The sharpness of the transition from no solution to a solution depends on the data and particular implementation. As C exceeds this value, the weight vector becomes less sparse.

2.7. Sparse hyperplane performance

Two metrics were devised to assist in evaluating the generalization performance of a sparse hyperplane estimated from data $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$ or $\{\mathcal{E}_n(\tilde{\mathbf{x}}_n, \Sigma_n), y_n\}_{n=1}^N$. Let $\Sigma_* = \rho\Sigma$ be the covariance matrix for an ellipsoidal data uncertainty model, and let $\mathcal{H}(\mathbf{w}_*, b_*)$ be the optimal (nominal or robust) hyperplane.

The "ordinary rule" for classifying a data point $\mathbf{x}$ is as follows. If $\mathbf{w}_*^T\mathbf{x} > b_*$, $\mathbf{x}$ is assigned to the +1 class. If $\mathbf{w}_*^T\mathbf{x} < b_*$, $\mathbf{x}$ is identified with the -1 class. An ordinary error occurs when the class predicted by the hyperplane differs from the known class of the data point.

The "worst case rule" determines whether an ellipsoid with center $\mathbf{x}$ intersects the hyperplane. Some allowable values of $\mathbf{x}$ will be classified incorrectly if $|\mathbf{w}_*^T\mathbf{x} + b_*| < \|\Sigma_*^{1/2}\mathbf{w}_*^T\|_2$, i.e., $m < r$ in Fig. 1. A worst case error occurs if the data point has an ordinary error (lies on the wrong side of the decision boundary) or if the ellipsoid and hyperplane overlap (some permissible values lie on the wrong side of the decision boundary).

The number of relevant features at noise ρ is the number of nonzero elements of the optimal weight vector $\mathbf{w}_*$, $\mathcal{F}(\rho) = \text{cardinality}\{p : w_*^p \neq 0\}, \quad p = 1, \ldots, P$.

3. RESULTS

The performance of nominal and robust sparse hyperplanes was investigated using an illustrative, real-world, two-class, transcriptional profiling dataset (Section 2.1). The breast tumor sample profiles represent transcript abundances in $N = 22$ samples, $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$. Each data point is a vector of log_2 transformed measurements, $\mathbf{x}_n = \{x_n^1, \ldots, x_n^P\}$, where $P = 3{,}226$ and x_n^p is the observed value of feature p in data point n. The 7 BRCA1 samples are assigned to the class $y_n = +1$; the 15 BRCA2 and sporadic samples belong to the other class, $y_n = -1$.

3.1. Nominal sparse hyperplane performance

In the sparse hyperplane optimization problem, the regularization parameter C plays a key role in determining both classification performance and sparsity of the weight vector. Nominal sparse hyperplanes

Table 1. The Performance of Nominal Liknon Linear Classifiers Learned Using Transcriptional Profiling Data (3,226 Genes, 22 Samples) and Different Settings of the Regularization Parameter C (the Noise Level Parameter ρ Was Zero)[a]

C	*Errors*	$\mathcal{F}(\rho = 0.0)$
0.1082275	22	0
0.1082276	2	7
0.109	2	7
0.11	2	7
0.1125	2	8
0.125	1	10
0.15	1	14
0.2	1	14
0.3	0	17

[a]For each C value examined, the table lists the number of ordinary errors out of 22 (Errors) and the number of relevant features out of 3,226 ($\mathcal{F}(\rho = 0.0)$).

were computed using the profiling data for a range of C values. For each C-specific nominal Liknon linear classifier, the ordinary error of the data and the number of relevant features (nonzero elements of the weight vector) were determined. The results are given in Table 1.

The solution to the optimization problem depends upon C. There is a sharp transition between no solution (all elements of the weight vector are zero) and a good solution. At $C = 0.1082275$, there is no solution so every point is classified incorrectly. At $C = 0.1082276$, there are two ordinary errors and seven relevant features. As C increases, the sparse hyperplane makes fewer errors, and the weight vector becomes less sparse. A value of $C = 0.3$ yields a nominal sparse hyperplane that generalizes well (zero ordinary error) and identifies a small number of relevant features (17 genes).

3.2. Robust sparse hyperplane performance

Robust sparse hyperplanes were computed using ellipsoidal data uncertainty models at a range of C and ρ values. Two types of models were examined, (i) the shape of the uncertainty was assumed to be the same for all data points, i.e., class-independent covariance matrices were computed using all 22 data points ($\Sigma_{\pm}$), and (ii) the shape of the uncertainty was assumed to be different for each category, i.e., class-independent covariance matrices were computed using 7 BRCA1 data points (Σ_{+}) and 15 BRCA1/sporadic data points (Σ_{-}). In each case, these matrices were diagonal and feature dependant (Equation 12). For each $C-$ and ρ-specific robust Liknon linear classifier, the ordinary error, worst case error, and number of relevant features were determined. The results are given in Table 2.

The sensitivity of sparse hyperplanes to misclassified data points can be adjusted via C. Larger values attempt to achieve full linear separability (potentially overfitting the data) whereas smaller values may ignore some outliers. The two largest settings, $C = 0.2$ and 0.5, yield few relevant features ("$\mathcal{F}(\rho)$") and classifiers that generalize well (low "ordinary" and "worst case" error). This is consistent with results using nominal Liknon in which a nominal sparse hyperplane estimated using $C = 0.3$ has the best performance (Table 1).

For increasingly uncertain data, the hyperplane becomes less sparse (the weight vector $\mathbf{w}$ has more nonzero elements). As ρ becomes larger, the number of relevant features increases slightly, and both types of classification error become larger. Overall, robust sparse hyperplanes employ only 10–20 out of the 3,226 features (genes) to form the robust linear classifier. In some instances, a worst-case error is due to the hyperplane intersecting a data uncertainty ellipsoid rather than the point being on the wrong side of the decision boundary (see, for example, results for $C = 0.2$, $\rho = 0.1$, and $\Sigma_{\pm}$). In terms of worst case error, a $C = 0.5$ nominal Liknon hyperplane is robust at noise levels $\rho \leq 0.3$.

TABLE 2. THE PERFORMANCE OF ROBUST AND NOMINAL LIKNON LINEAR CLASSIFIERS COMPUTED USING TRANSCRIPTIONAL PROFILING DATA AND DIFFERENT SETTINGS OF THE REGULARIZATION PARAMETER C AND NOISE LEVEL PARAMETER ρ[a]

		Covariance matrix: $\Sigma_{\pm}$				*Covariance matrices:* Σ_+, Σ_-			
		Robust			*Nominal*	*Robust*			*Nominal*
C	ρ	$\mathcal{F}(\rho)$	*Ordinary*	*Worst case*	*Worst case*	$\mathcal{F}(\rho)$	*Ordinary*	*Worst case*	*Worst case*
0.109	0	7	2	NA	NA	7	2	NA	NA
	0.001	8	2	2(0)	2(0)	0	x	x	2(0)
	0.01	8	2	2(0)	2(0)	0	x	x	2(0)
	0.1	7	2	3(1)	3(1)	0	x	x	2(0)
	0.2	8	2	4(2)	4(2)	0	x	x	2(0)
	0.3	8	2	7(7)	7(7)	0	x	x	3(1)
	0.5	0	x	x(0)	21(21)	0	x	x	3(1)
	0.7	6[b]	x	x(0)	21(21)	0	x	x	4(2)
0.125	0	10	1	NA	NA	10	1	NA	NA
	0.001	11	1	1(0)	1(0)	10	1	1(0)	1(0)
	0.01	11	1	1(0)	1(0)	10	1	1(0)	1(0)
	0.1	11	1	1(0)	1(0)	12	1	1(0)	1(0)
	0.2	10	1	2(1)	2(0)	12	1	1(0)	1(0)
	0.3	12	1	2(2)	2(0)	16	0	1(0)	2(1)
	0.5	16	1	13(13)	19(19)	16	0	1(0)	2(1)
	0.7	17	2	16(16)	21(21)	21	0	4(4)	2(1)
0.2	0	14	1	NA	NA	14	1	NA	NA
	0.001	14	1	1(0)	1(0)	16	0	0(0)	1(0)
	0.01	14	1	1(0)	1(0)	17	0	0(0)	1(0)
	0.1	14	1	1(1)	1(0)	20	0	0(0)	1(0)
	0.2	15	1	1(1)	1(0)	15	0	0(0)	1(0)
	0.3	16	1	2(2)	2(2)	19	0	1(1)	1(0)
	0.5	26	0	10(10)	20(20)	19	0	1(1)	1(1)
	0.7	42	0	10(10)	22(22)	19	0	1(1)	2(2)
0.5	0	17	0	NA	NA	17	0	NA	NA
	0.001	17	0	0(0)	0(0)	17	0	0(0)	0(0)
	0.01	17	0	0(0)	0(0)	17	0	0(0)	0(0)
	0.1	19	0	0(0)	0(0)	17	0	0(0)	0(0)
	0.2	22	0	0(0)	0(0)	18	0	0(0)	0(0)
	0.3	27	0	0(0)	0(0)	20	0	0(0)	0(0)
	0.5	33	0	7(7)	21(21)	18	0	0(0)	0(0)
	0.7	52	0	8(8)	22(22)	18	0	0(0)	0(0)

[a]For robust sparse hyperplanes, the shape of the ellipsoidal data uncertainty was the same for all data points (class-independent covariance matrix, $\Sigma_{\pm}$) or different for each category (class-dependent covariance matrices, Σ_+ and Σ_-). For nominal sparse hyperplanes, the results are for the nominal classifier found using $\rho = 0$. The abbreviations and symbols are "Robust," robust LIKNON; "Nominal," nominal LIKNON ($\rho = 0$); "$\mathcal{F}(\rho)$," number of relevant features out of 3,226; "Ordinary," ordinary error out of 22; "Worst case," worst case error out of 22 with the number in parenthesis indicating the number points where the ellipsoid and hyperplane intersect; "x," no solution could be found; "NA," calculation of worst case error requires that the data are associated with some uncertainty, i.e., $\rho > 0$.

[b]A six gene solution in which the corresponding elements in the weight vector had values just above the $1E - 8$ threshold used to assign zero-elements. No solutions were found for $C < 0.109$.

Ellipsoidal data uncertainty models utilizing class-dependent covariance matrices (Σ_+, Σ_-) are preferable to ones using a class-independent matrix ($\Sigma_\pm$). When the shape of the uncertainty associated with points in each category is modeled separately, the robust sparse hyperplane specifies fewer features and has smaller errors than one where the category is ignored. This difference is most marked at high noise levels, $\rho > 0.3$.

Robust sparse hyperplanes are more resilient to data uncertainty than are their nominal counterpart. Robust LIKNON hyperplanes have fewer ordinary and worst-case errors than nominal LIKNON hyperplanes (see, for example, results for $C = 0.2$, $\rho = 0$–0.7 for both types of covariance matrices).

3.3. Number of relevant features $\mathcal{F}(\rho)$

There exist many small feature subsets that are equally good at distinguishing classes; i.e., there is no single "best" set of discriminatory features (see also Chow *et al.* [2001]). For the dataset examined here, a robust sparse hyperplane estimated using regularization parameter $C = 0.5$, noise level $\rho = 0.0$–0.7, and diagonal, feature-dependent, class-dependent covariance matrices yields a linear classifier with 17–20 genes, and zero ordinary and worst case error. The benefit of a multiplicity of accurate classifiers is the ability to develop clinical tests, each based on a different set of tens of genes, that generalize well to new patients. The disadvantage is that experimentalists interested in molecular mechanisms and evaluating targets for therapeutic intervention would prefer few genes.

3.4. Formulation and numerical issues

The hyperplane parameter values found are sensitive to the SOCP formulation that is solved. Table 2 shows that at the regularization parameter threshold, $C = 0.109$, a solution is found for the formulation based on a class-independent covariance matrix but there is no solution when class-dependent matrices are used. Numerical issues also lead to a slight variation in the number of relevant features. Once a solver has found an optimal weight vector, the "nonzero" elements of $\mathbf{w}$ are determined according to whether w_p exceeds a "small" threshold. This work employed a cutoff of $1E-8$, so any value below this is set to zero.

Table 3 lists relevant features for two specific nominal and robust LIKNON linear classifiers. For the robust LIKNON 18 and 20 gene solutions listed in Table 2, the weights of the additional genes outside the 17 in Table 3 are very small.

3.5. Implementation and runtimes

The computational experiments described here used nominal and robust sparse hyperplanes implemented using Matlab and run on a 466 MHz DEC Alpha computer in a typical university network environment. For the core algorithm, runtimes were 13 seconds for nominal LIKNON (using the built-in linear solver linprog), 60 seconds for robust LIKNON (using SeDuMi) with a class-independent covariance matrix, and 75 seconds for robust LIKNON (using `SeDuMi`) with class-dependent covariance matrices. An additional 15 seconds were required for starting up Matlab, reading in files, and setting up the problem. The memory requirements were 82 and 54 MBytes for robust and nominal LIKNON, respectively.

4. DISCUSSION

Two-class high-dimensional data arise in many domains. Key statistical tasks are learning a classifier that generalizes well and identifying a small number of features able to distinguish the classes. Previously, it was shown that a sparse hyperplane can address these tasks simultaneously because the parameters of such a model, the weight vector and offset, define a linear decision surface and the nonzero elements of the weight vector specify the discriminatory features. This work (i) proposes robust sparse hyperplanes as a classification and relevant feature identification method that is resilient to uncertainty in the data points, (ii) formulates the optimization problem for estimating a model from data as an SOCP, and (iii) demonstrates the potential of a specific implementation, robust LIKNON, on an illustrative, real-world dataset. Because

TABLE 3. GENES DISCRIMINATING 7 BRCA1 BREAST TUMOR SAMPLES FROM 15 BRCA2/SPORADIC SAMPLES

Robust LIKNON: $C = 0.5$; $\rho = 0.001$; diagonal, feature-dependent covariance matrices Σ_+, Σ_-

Feature	*Weight*	*IMAGE*	*Annotation*
179	−0.146735	809627	Nuclear receptor interacting protein 1
297	−0.010796	246786	Human orphan G protein-coupled receptor (RDC1) mRNA, partial cds
336	−0.447493	823940	Transducer of ERBB2, 1[a]
435	0.160054	45542	Human insulin-like growth factor binding protein 5 (IGFBP5) mRNA
478	0.108011	309032	Human cleavage and polyadenylation specificity factor mRNA, complete cds
739	−0.010240	214068	GATA-binding protein 3[a]
1008	−0.278422	897781	Keratin 8
1697	−0.033875	247233	ESTs
1934	0.025315	66977	ESTs, highly similar to CGI-103 protein (H. sapiens)
2226	−0.001733	282980	ESTs
2272	0.104417	309583	ESTs[a]
2632	−0.017060	40111	Transcription factor 8 (represses interleukin 2 expression)
2633	−0.041071	40151	Apolipoprotein D
2890	0.033002	26167	Fas (TNFRSF6)-associated via death domain
2893	0.154681	32790	MutS (E. coli) homolog 2 (colon cancer, nonpolyposis type 1)
3080	0.062061	280768	Transmembrane 4 superfamily member 1[a]
3199	0.155637	375635	Transcription factor 12 (HTF4, helix-loop-helix transcription factors 4)

Nominal LIKNON: $C = 0.109$

Feature	*Weight*	*IMAGE*	*Annotation*
336	−0.170155	823940	Transducer of ERBB2, 1[a]
739	−0.162721	214068	GATA−binding protein 3[a]
991	0.022604	46916	Matrix metalloproteinase 16 (membrane-inserted)[a]
1482	0.356199	839736	Crystallin, alpha B
1859	0.049336	307843	ESTs
2272	0.061724	309583	ESTs[a]
3080	0.060753	280768	Transmembrane 4 superfamily member 1[a]

[a]Relevant features common to the nominal and robust LIKNON linear classifiers.

of practical limitations imposed by extant solvers, the computational cost of the SOCP is reduced by introducing similar shape matrices for sets of data points. In the future, however, such restrictions should be removed yielding enhanced models with potentially better performance. Although the robust sparse hyperplanes developed here were motivated by a need to analyze transcriptional profiling data, they can be applied to (noisy) two-class, high-dimensional data from other areas.

Recently, Kim *et al.* (2002) proposed an alternative strategy for determining a linear classifier that is robust to noise in transcriptional profiling data. Given a set of genes (features), the hyperplane is computed using an analytic spherical noise model. Sets of 1–3 genes are found via an exhaustive search that enumerates all combinations. This strategy becomes intractable for larger sets so a guided random walk is used for four or more genes. For the data set examined here, 140 hours on a supercomputer cluster was needed to identify at least 11 pairs of genes that separate the data. Estimating a hyperplane for a set of genes is fast so it is likely that most of the time was spent discovering gene sets. In contrast, the robust sparse hyperplanes proposed here permit use of a more complex data uncertainty model (ellipsoids with different shapes), can determine discriminatory genes at the same time as learning the hyperplane parameters, and are computationally less demanding.

On the same breast cancer transcript profiles, robust LIKNON finds a sparse solution involving 7–10 genes that separates the data, whereas the approach of Kim *et al.* finds a sparser solution involving 2 genes. Recall that minimizing the l_0 norm of the weight vector (subject to low classification error), $\|\mathbf{w}\|_0$, would yield a sparse hyperplane. This is an NP-hard problem so the tractable convex approximation exploited here (and

by others) is to minimize the l_1 norm, $\|\mathbf{w}\|_1$. Thus, although both small gene sets generalize equally well, the minimal l_1 solution is not a minimal l_0 solution (2 genes). This discrepancy will probably arise in other data sets but the magnitude will differ. We are not aware of any formal analysis which addresses the difference, if any, between minimal l_0 and l_1 solutions when the test error is identical.

It is unlikely that the sparse hyperplane $min\|\mathbf{w}\|_1$ problem can be modified to yield a number of nonzero elements close to the $min\|\mathbf{w}\|_0$ solution. First, sparsity is sensitive to many facets of the data (quality, precision, preprocessing, noise level ρ), the free-parameter setting used to estimate a hyperplane (regularization parameter C), and implementation issues (32- or 64-bit computer, particular solver, and so on). The l_1-based cost function minimizes a linear combination of the weight vector elements, w^p, so their absolute magnitudes are important. In this simple unweighted sum, if w^p is large, the corresponding feature in the data, x^p, is small, and vice versa. This cost function has no direct control over the number of genes in a solution. Numerically, a 20-gene ("mediocre" scoring) solution where $w^p = 0.01$ is "better" than a 2-gene ("low" scoring) solution where $w^p = 1$, i.e., $(20 * 0.01 = 0.2) < (2 * 1 = 2)$. Biologically, however, the 2-gene solution is probably desirable. Smaller w^p—less sparse vectors—are more likely because there are many features with sufficiently large x^p values that can form solutions with small w^p values compared to ones with small x^p values that would require large w^p values. Clearly, the data preprocessing step can influence the final solution so any transformation applied should lead to larger x^p values for characteristics a user wishes to emphasize.

Despite the limitations discussed above, robust sparse hyperplanes hold potential as a technique for analyzing molecular profiles in particular and two-class high-dimensional data in general. Future directions include exploring different generative models for the data and more general covariance structures (Boyd and Vandenberghe, 2003). It may become necessary to formulate robust sparse classifiers in which classes are separated by nonlinear decision boundaries, a formulation which can be achieved via the use of Mercer kernels.

APPENDIX

Statistical interpretation of the robust LP

The fundamental classification problem can be viewed from a statistical perspective (Boyd and Vandenberghe, 2003). The basic optimization problem for learning a sparse hyperplane (1) has two terms and various constraints. The first term pertains to the weight vector, $\|\mathbf{w}\|_1$, and a second is designed to account for possible misclassifications in the data, $C\sum_{n=1}^{N}\xi_n$. These two terms are the same in all formulations and so are not relevant to subsequent discussions. The focus here is the noise models in the constraints.

Consider the situation in which each deterministic constraint, $y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 0$, should hold with a probability or confidence exceeding η,

$$\begin{aligned} \min_{\mathbf{w}} \quad & \|\mathbf{w}\|_1 \\ \text{subject to} \quad & \text{Prob}(y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 0) \geq \eta \\ & \eta \geq 0, \quad n = 1, \ldots, N. \end{aligned} \tag{18}$$

Higher values of η denote more a stringent requirement that the point $\mathbf{x}_n$ belong to the correct half-space.

The probability constraint in (18) can be expressed as a second-order cone constraint identical in character to that based on the use of ellipsoidal data uncertainty models. The two formulations derived below make different assumptions about the distribution of the data $\mathbf{x}$: Gaussian or distribution-free. Both scenarios presume that a data point, $\mathbf{x} \in \mathbb{R}^P$, is a random variable whose distribution is specified by a mean, $\tilde{\mu} \in \mathbb{R}^P$, and covariance matrix, $\Sigma \in \mathbb{R}^{P\times P}$.

Probability constraint for Gaussian random variable

Let $\eta \geq 0.5$, and assume that the random variable is distributed according to a Gaussian with mean $\tilde{\mu}$ and variance $\Sigma^2 \in \mathbb{R}$, $\mathbf{x} \sim \mathcal{N}(\tilde{\mu}, \Sigma^2)$. Let $\mu = -(\mathbf{w}^T\mathbf{x})$, $\tilde{\mu} = -(\mathbf{w}^T\tilde{\mathbf{x}})$, and $\Sigma = (\mathbf{w}^T\Sigma\mathbf{w})^{1/2} \equiv \|\Sigma^{1/2}\mathbf{w}\|_2$.

Substituting and rearranging, the probability constraint in (18) can be written as

$$\text{Prob}\left(\frac{\mu-\tilde{\mu}}{\Sigma} \leq \frac{b-\tilde{\mu}}{\Sigma}\right) \geq \eta. \tag{19}$$

By definition, $(\mu-\tilde{\mu})/\Sigma$ is a zero-mean unit variance Gaussian random variable so the probability (19) becomes

$$\text{Prob}\left(\frac{\mu-\tilde{\mu}}{\Sigma} \leq \frac{b-\tilde{\mu}}{\Sigma}\right) = \Phi\left(\frac{b-\tilde{\mu}}{\Sigma}\right) \geq \eta$$
$$\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{t^2}{2}}dt. \tag{20}$$

Thus, the probability constraint in (18) can be expressed as

$$\frac{b-\tilde{\mu}}{\Sigma} \geq \Phi^{-1}(\eta), \quad \text{or equivalently,} \quad \tilde{\mu}+\Phi^{-1}(\eta)\Sigma \leq b. \tag{21}$$

Since $\eta \geq 0.5$, $\Phi^{-1}(\eta) \geq 0$ so the constraint is a second-order constraint and can be written as

$$y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq \Phi^{-1}(\eta)\|\Sigma_n^{1/2}\mathbf{w}\|_2$$
$$\eta \geq 0.5, \quad n = 1, \ldots, N. \tag{22}$$

Given that η is specified a priori, $\Phi^{-1}(\eta)$ is simply a constant which scales the covariance matrix Σ and is thus identical to the ellipsoid constraint in (5). Thus, an LP in which the data are treated as Gaussian random variables yields an SOCP constraint (22) that, up to a multiplicative factor, is equivalent to one based on an ellipsoidal data uncertainty model (5).

Probability constraint for the distribution-free setting

Transcript profiling data seldom follow a Gaussian distribution. In situations when only second-order statistics are available, the multivariate Chebyshev bound and a worst-case setting can be used to express the probability constraint (18). Assume that only the first two moments of a random variable, $\mathbf{x}$, are known. Let $\mathbf{x} \sim (\tilde{\mu}, \Sigma)$ denote all possible distributions with mean $\tilde{\mu}$ and covariance matrix Σ. An existing theorem (Marshall and Olkin, 1960; Popescu and Bertsimas, 2001) states that the supremum of the probability that a random vector takes a value in an arbitrary closed convex set $\mathcal{S}$ is

$$\sup_{\mathbf{x}\sim(\tilde{\mu},\Sigma)} \text{Prob}(\mathbf{x} \in \mathcal{S}) = \frac{1}{1+d^2}$$
$$d^2 = \inf_{\alpha\in S}(\alpha-\tilde{\mu})^T\Sigma(\alpha-\tilde{\mu}). \tag{23}$$

Using the above observation and the fact that a half-space produced by a hyperplane is a closed convex set, previous work (Lanckriet *et al.*, 2002) has shown that the probability constraint in (18) can be expressed as a second-order cone constraint,

$$\mathbf{w}^T\tilde{\mu} + b \geq \sqrt{\frac{\eta}{(1-\eta)}}\sqrt{\mathbf{w}^T\Sigma\mathbf{w}}. \tag{24}$$

Hence, the original problem constraint in (18) can be rewritten as

$$y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq \kappa(\eta)\|\Sigma_n^{1/2}\mathbf{w}\|_2$$
$$\eta \geq 0, \quad n = 1, \ldots, N \tag{25}$$

where $\kappa(\eta) = \sqrt{\frac{\eta}{1-\eta}}$.

Given that η is specified a priori, $\kappa(\eta)$ is simply a constant which scales the covariance matrix Σ and, again, has the same form as the above Gaussian case. Thus, an LP in which the data are treated as random variables for which only the first two moments are known yields an SOCP constraint (25) that is, up to a multiplicative factor, identical to one based on an ellipsoidal data uncertainty model (5), and one that assumes Gaussian random variables (22).

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation, Office of Naval Research, National Institute on Aging, National Institute of Environmental Health Sciences, U.S. Department of Energy, and California Breast Cancer Research Program.

REFERENCES

Amaldi, E., and Kann, V. 1998. On the approximability of minimizing non zero variables or unsatisfied relations in linear systems. *Theoret. Comput. Sci.* 209, 237–260.

Ben-Tal, A., El Ghaoui, L., and Nemirovskii, A. 2000. *Robust Semidefinite Programming*, Kluwer Academic, Waterloo, Canada.

Bennett, K., and Campbell, C. 2000. Support vector machines: Hype or hallelujah? *SIGKDD Explorations* 2, 1–13.

Bennett, K., and Demiriz, A. 1999. Semi-supervised support vector machines, *in Advances in Neural and Information Processing Systems*, vol. 11, 368–374, MIT Press, Cambridge, MA.

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

Bhattacharyya, C., Grate, L., Rizki, A., Radisky, D., Molina, F., Jordan, M., Bissell, M., and Mian, I. 2003. Simultaneous classification and relevant feature identification in high-dimensional spaces: Application to molecular profiling data. *Signal Processing* 83, 729–743.

Boyd, S., and Vandenberghe, L. 2003. *Convex Optimization*, Cambridge University Press, Cambridge, UK.

Chow, M., Moler, E., and Mian, I. 2001. Identifying marker genes in transcription profile data using a mixture of feature relevance experts. *Physiological Genomics* 5, 99–111.

Cristianini, N., and Shawe-Taylor, J. 2000. *Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, UK.

Dhanasekaran, S., Barrette, T., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K., Rubin, M., and Chinnaiyan, A. 2001. Delineation of prognostic biomarkers in prostate cancer. *Nature* 432, 822–826.

Donoho, D., and Huo, X. 1999. Uncertainty principles and ideal atomic decomposition. Technical report, Statistics Department, Stanford University. The report is available at *www-stat.stanford.edu/ donoho/reports.html*.

Garber, M., Troyanskaya, O., Schluens, K., Petersen, S., Thaesler, Z., Pacyana-Gengelbach, M., van de Rijn, M., Rosen, G., Perou, C., Whyte, R., Altman, R., Brown, P., Botstein, D., and Petersen, I. 2001. Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl. Acad. Sci.* 98, 13784–13789.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfeld, C., and Lander, E. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537.

Graepel, T., Herbrich, R. Schölkopf, B., Smola, A., Bartlett, P., Müller, K., Obermayer, K., and Williamson, R. 1999. Classification on proximity data with lp-machines. *Proc. 9th Int. Conf. on Artificial Neural Networks* 470, 304–309.

Grate, L., Bhattacharyya, C., Jordan, M., and Mian, I. 2002. Simultaneous relevant feature identification and classification in high-dimensional spaces. *Workshop on Algorithms in Bioinformatics (WABI 2002)*, 1–9, Springer, Rome, Italy.

Grate, L., Bhattacharyya, C., Jordan, M., and Mian, I. 2003. Integrated analysis of transcript profiling and protein sequence data. *Mechanisms of Ageing and Development* 124, 109–114.

Hastie, T., Tibshirani, R., and Friedman, J. 2000. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, New York.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S.,

Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O.-P., Borg, A., and Trent, J. 2001. Gene-expression profiles in hereditary breast cancer. *New England J. Med.* 344, 539–548.

Khan, J., Wei, J., Ringner, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, C., and Meltzer, P. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* 7, 673–679.

Kim, S., Dougherty, E., Barrera, J., Chen, Y., Bittner, M., and Trent, J. 2002. Strong feature sets form small samples. *J. Comp. Biol.* 9, 127–146.

Lanckriet, G., El Ghaoui, L., Bhattacharyya, C., and Jordan, M. 2002. A robust minimax approach to classification. *J. Machine Learning Res.* 3, 555–582.

Liotta, L., Kohn, E., and Perticoin, E. 2001. Clinical proteomics: Personalized molecular medicine. *JAMA* 14, 2211–2214.

Marshall, A., and Olkin, I. 1960. Multivariate Chebyshev inequalities. *Ann. Math. Statist.* 31, 1001–1014.

Notterman, D., Alon, U., Sierk, A., and Levine, A. 2001. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res.* 61, 3124–3130.

Novak, J., Sladek, R., and Hudson, T. 2002. Characterization of variability in large-scale gene expression data: Implications for study design. *Genomics* 79, 104–113.

Popescu, I., and Bertsimas, D. 2001. Optimal inequalities in probability theory. Technical report TB 62, INSEAD.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., Poggio, T., Gerald, W., Loda, M., Lander, E., and Golub, T. 2001. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci.* 98, 15149–15154.

Smola, A., Frieß, T., and Schölkopf, B. 1999. Semiparametric support vector and linear programming machines, *in Neural and Information Processing Systems*, vol. 11, MIT Press, Cambridge, MA.

Sørlie, T., Perou, C., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M., van de Rijn, M., Jeffrey, S., Thorsen, T., Quist, H., Matese, J., Brown, P., Botstein, D., Lønning, P., and Børresen-Dale, A.-L. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.* 98, 10869–10874.

Sturm, J. 1999. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software* 11, 625–653.

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

Weston, J., Elisseeff, A., Schölkopf, B., and Tipping, M. 2003. Use of the zero-norm with linear models and kernel methods. *J. Machine Learning Res.* 3, 1439–1461.

Address correspondence to:
C. Bhattacharyya
Department of Computer Science and Automation
Indian Institute of Science
Bangalore 560012
Karnataka, India

E-mail: chiru@csa.iisc.ernet.in