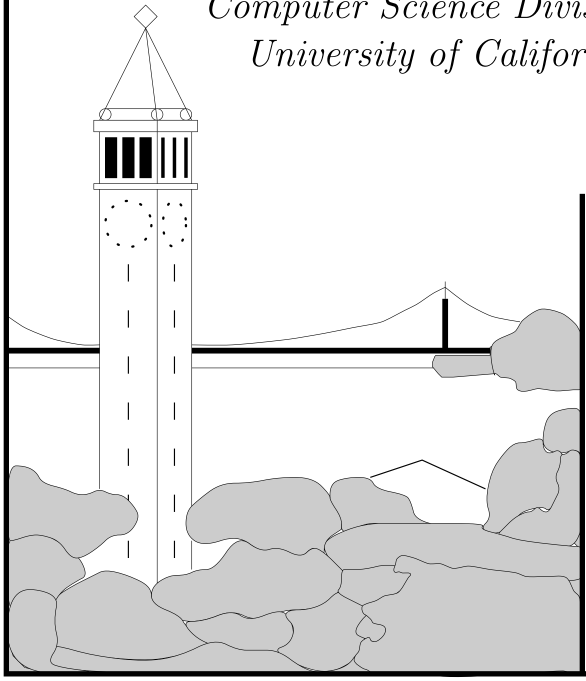
A Robust Minimax Approach to Classification

Gert R.G. Lanckriet
gert@cs.berkeley.edu
Department of Electrical Engineering and Computer Science
University of California, Berkeley, CA 94720, USA

Laurent El Ghaoui
elghaoui@eecs.berkeley.edu
Department of Electrical Engineering and Computer Science
University of California, Berkeley, CA 94720, USA

Chiranjib Bhattacharyya
chiru@csa.iisc.ernet.in
Department of Computer Science and Automation
Indian Institute of Science, Bangalore 560012, Karnataka, India

Michael I. Jordan
jordan@cs.berkeley.edu
Computer Science Division and Department of Statistics
University of California, Berkeley, CA 94720, USA



Report No. UCB/CSD-02-1218

December, 2002

Computer Science Division (EECS)
University of California
Berkeley, California 94720

A Robust Minimax Approach to Classification

Gert R.G. Lanckriet
gert@cs.berkeley.edu
Department of Electrical Engineering and Computer Science
University of California, Berkeley, CA 94720, USA

Laurent El Ghaoui
elghaoui@eecs.berkeley.edu
Department of Electrical Engineering and Computer Science
University of California, Berkeley, CA 94720, USA

Chiranjib Bhattacharyya
chiru@csa.iisc.ernet.in
Department of Computer Science and Automation
Indian Institute of Science, Bangalore 560012, Karnataka, India

Michael I. Jordan
jordan@cs.berkeley.edu
Computer Science Division and Department of Statistics
University of California, Berkeley, CA 94720, USA

December, 2002

Abstract

When constructing a classifier, the probability of correct classification of future data points should be maximized. We consider a binary classification problem where the mean and covariance matrix of each class are assumed to be known. No further assumptions are made with respect to the class-conditional distributions. Misclassification probabilities are then controlled in a worst-case setting: that is, under all possible choices of class-conditional densities with given mean and covariance matrix, we *mini*mize the worst-case (*max*imum) probability of misclassification of future data points. For a linear decision boundary, this desideratum is translated in a very direct way into a (convex) second order cone optimization problem, with complexity similar to a support vector machine problem. The minimax problem can be interpreted geometrically as minimizing the maximum of the Mahalanobis distances to the two classes. We address the issue of robustness with respect to estimation errors (in the means and covariances of the classes) via a simple modification of the input data. We also show how to exploit Mercer kernels in this setting to obtain nonlinear decision boundaries, yielding a classifier which proves to be competitive with current methods, including support vector machines. An important feature of this method is that a worst-case bound on the probability of misclassification of future data is always obtained explicitly.

Keywords: classification, kernel methods, convex optimization, second order cone programming

1 Introduction

Consider the problem of choosing a linear discriminant by minimizing the probabilities that data vectors fall on the wrong side of the boundary. One way to attempt to achieve this is via a generative approach in which one makes distributional assumptions about the class-conditional densities and thereby estimates and controls the relevant probabilities. The need to make distributional assumptions, however, casts doubt on the generality and validity of such an approach, and in discriminative solutions to classification problems it is common to attempt to dispense with class-conditional densities entirely.

Rather than avoiding any reference to class-conditional densities, it might be useful to attempt to control misclassification probabilities in a worst-case setting; that is, under all possible choices of class-conditional densities with a given mean and covariance matrix, minimize the worst-case (maximum) probability of misclassification of future data points. Such a minimax approach could be viewed as providing an alternative justification for discriminative approaches. In this paper we show how such a minimax program can be carried out in the setting of binary classification. Our approach involves exploiting the following powerful theorem due to Marshall and Olkin (1960), which has recently been put in the light of recent convex optimization techniques by Popescu and Bertsimas (2001):

$$\sup_{\mathbf{y} \sim (\bar{\mathbf{y}}, \mathbf{\Sigma}_{\mathbf{y}})} \mathbf{Pr}\{\mathbf{y} \in \mathcal{S}\} = \frac{1}{1+d^2} \ , \text{ with } \quad d^2 = \inf_{\mathbf{y} \in \mathcal{S}} \, (\mathbf{y} - \bar{\mathbf{y}})^T {\mathbf{\Sigma}_{\mathbf{y}}}^{-1} (\mathbf{y} - \bar{\mathbf{y}}),$$

where $\mathbf{y}$ is a random vector, $\mathcal{S}$ is a given convex set, and where the supremum is taken over all distributions for $\mathbf{y}$ having mean $\bar{\mathbf{y}}$ and covariance matrix $\mathbf{\Sigma}_{\mathbf{y}}$ (assumed to be positive definite for simplicity). This theorem provides us with the ability to bound the probability of misclassifying a point, without making Gaussian or other specific distributional assumptions. We will show how to exploit this ability in the design of linear classifiers, for which the set $\mathcal{S}$ is a half-space. The usefulness of this theorem will also be illustrated in the case of unsupervised learning, for the single class problem.

One of the appealing features of this formulation is that one obtains an explicit upper bound on the probability of misclassification of future data: $1/(1+d^2)$. We will also show that the learning problem can be interpreted geometrically, as yielding that hyperplane that minimizes the maximum of the Mahalanobis distances from the class means to the hyperplane.

Another feature of this formulation is the possibility to study directly the effect of using plug-in estimates for the means and covariance matrices rather then their real, but unknown, values. We will assume the real mean and covariance are unknown, but bounded in a convex region and show how this affects the resulting classifier: it amounts to incorporating a regularization term in the estimator and results in an increase in the upper bound on the probability of misclassification of future data.

A third important feature of this approach is that, as in linear discriminant analysis (Mika et al., 1999), it is possible to generalize the basic methodology to allow nonlinear decision boundaries via the use of Mercer kernels. The resulting nonlinear classifiers are competitive with existing classifiers, including support vector machines.

The paper is organized as follows: in Section 2 we present the minimax formulation for linear classifiers and our main results. In Section 3, we show how a simple modification of our method makes it possible to treat estimation errors in the mean and covariance matrices within the minimax formulation. The single class case is studied in Section 4. In Section 5 we deal with kernelizing the method. Finally, we present empirical results in Section 6. The results in Sections 2, 5 and 6 have been presented earlier (Lanckriet et al., 2002a), and are expanded significantly in the current paper, while the results in Sections 3 and 4 are new. Matlab code to build and evaluate the different types of classifiers can be downloaded from *http://robotics.eecs.berkeley.edu/~gert/*.

2 Minimax Probabilistic Decision Hyperplane

In this section, the minimax formulation for linear classifiers is presented. After defining the problem in Section 2.1, we state the main result in Section 2.2 and propose an optimization algorithm in Section 2.3. In Section 2.4 and 2.5, we show how the optimization problem can be interpreted in a geometric way. Section 2.6 addresses the effect of making Gaussian assumptions and the link with Fisher discriminant analysis is discussed in Section 2.7.

2.1 Problem Definition

Let $\mathbf{x}$ and $\mathbf{y}$ denote random vectors in a binary classification problem, modelling data from each of two classes, with means and covariance matrices given by $(\bar{\mathbf{x}}, \mathbf{\Sigma_x})$ and $(\bar{\mathbf{y}}, \mathbf{\Sigma_y})$, respectively, with $\mathbf{x}, \bar{\mathbf{x}}, \mathbf{y}, \bar{\mathbf{y}} \in \mathbb{R}^n$, and $\mathbf{\Sigma_x}, \mathbf{\Sigma_y} \in \mathbb{R}^{n \times n}$ both symmetric and positive semidefinite. With a slight abuse of notation, we let $\mathbf{x}$ and $\mathbf{y}$ denote the classes.

We wish to determine a hyperplane $\mathcal{H}(\mathbf{a}, b) = \{\mathbf{z} \mid \mathbf{a}^T\mathbf{z} = b\}$, where $\mathbf{a} \in \mathbb{R}^n \backslash \{0\}$ and $b \in \mathbb{R}$, which separates the two classes of points with maximal probability with respect to all distributions having these means and covariance matrices. This is expressed as

$$\max_{\alpha, \mathbf{a} \neq 0, b} \alpha \quad \text{s.t.} \quad \begin{aligned} &\inf_{\mathbf{x} \sim (\bar{\mathbf{x}}, \mathbf{\Sigma_x})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x} \geq b\} \geq \alpha \\ &\inf_{\mathbf{y} \sim (\bar{\mathbf{y}}, \mathbf{\Sigma_y})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \alpha, \end{aligned} \tag{1}$$

where the notation $\mathbf{x} \sim (\bar{\mathbf{x}}, \mathbf{\Sigma_x})$ refers to the class of distributions that have prescribed mean $\bar{\mathbf{x}}$ and covariance $\mathbf{\Sigma_x}$, but are otherwise arbitrary; likewise for $\mathbf{y}$.

Future points $\mathbf{z}$ for which $\mathbf{a}^T\mathbf{z} \geq b$ are then classified as belonging to the class associated with $\mathbf{x}$, otherwise they are classified as belonging to the class associated with $\mathbf{y}$. In formulation (1) the term $1 - \alpha$ is the worst-case (maximum) misclassification probability, and our classifier minimizes this maximum probability.

For simplicity, we will assume that both $\mathbf{\Sigma_x}$ and $\mathbf{\Sigma_y}$ are positive definite; our results can be extended to the general positive semidefinite case, using the appropriate extension of Marshall and Olkin's result, as discussed by Popescu and Bertsimas (2001). However, we do not handle this case here, since in practice we always add a regularization term to the covariance matrices. A complete discussion of the choice of the regularization parameter is given in Section 3.

2.2 Main Result

We begin with a lemma that will be used in the sequel.

Lemma 1 *Using the notation of Section 2.1, with* $\bar{\mathbf{y}}$*,* $\mathbf{\Sigma_y}$ *positive definite,* $\mathbf{a} \neq 0$*,* b *given, such that* $\mathbf{a}^T\bar{\mathbf{y}} \leq b$ *and* $\alpha \in [0,\, 1)$*, the condition*

$$\inf_{\mathbf{y} \sim (\bar{\mathbf{y}}, \mathbf{\Sigma_y})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \alpha \tag{2}$$

holds if and only if

$$b - \mathbf{a}^T\bar{\mathbf{y}} \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\mathbf{\Sigma_y}\mathbf{a}}, \tag{3}$$

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

Proof of lemma: In view of the result of Marshall and Olkin (1960) and using $\mathcal{S} = \{\mathbf{a}^T\mathbf{y} \geq b\}$, we obtain:

$$\sup_{\mathbf{y} \sim (\bar{\mathbf{y}}, \mathbf{\Sigma_y})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \geq b\} = \frac{1}{1+d^2}\ , \text{ with } \ d^2 = \inf_{\mathbf{a}^T\mathbf{y} \geq b} (\mathbf{y} - \bar{\mathbf{y}})^T {\mathbf{\Sigma_y}}^{-1}(\mathbf{y} - \bar{\mathbf{y}}). \tag{4}$$

The minimum distance d admits a simple closed-form expression (see Appendix A):

$$d^2 = \inf_{\mathbf{a}^T\mathbf{y} \geq b} (\mathbf{y} - \bar{\mathbf{y}})^T {\mathbf{\Sigma_y}}^{-1}(\mathbf{y} - \bar{\mathbf{y}}) = \frac{\max((b - \mathbf{a}^T\bar{\mathbf{y}}), 0)^2}{\mathbf{a}^T\mathbf{\Sigma_y}\mathbf{a}}. \tag{5}$$

The probability constraint (2) in this lemma is equivalent to $\sup_{\mathbf{y} \sim (\bar{\mathbf{y}}, \mathbf{\Sigma_y})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \geq b\} \leq 1 - \alpha$. Using (4), this becomes $1 - \alpha \geq 1/(1 + d^2)$ or $d^2 \geq \alpha/(1 - \alpha)$. Taking (5) into account, this is expressed as

$$\max((b - \mathbf{a}^T\bar{\mathbf{y}}), 0) \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\mathbf{\Sigma_y}\mathbf{a}} \qquad \text{where} \quad \kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}.$$

If $\mathbf{a}^T\bar{\mathbf{y}} \leq b$, we have $\max((b - \mathbf{a}^T\bar{\mathbf{y}}), 0) = b - \mathbf{a}^T\bar{\mathbf{y}}$ and this indeed reduces to

$$b - \mathbf{a}^T\bar{\mathbf{y}} \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\mathbf{\Sigma_y}\mathbf{a}}.$$

This concludes the proof. □

Our main result is stated below.

Theorem 2 *If* $\bar{\mathbf{x}} = \bar{\mathbf{y}}$*, then the minimax probability decision problem (1) does not have a meaningful solution: the optimal worst-case misclassification probability that we obtain is* $1-\alpha_* = 1$*. Otherwise, an optimal hyperplane* $\mathcal{H}(\mathbf{a}_*, b_*)$ *exists, and can be determined by solving the convex optimization problem:*

$$\kappa_*^{-1} := \min_{\mathbf{a}} \sqrt{\mathbf{a}^T\mathbf{\Sigma_x}\mathbf{a}} + \sqrt{\mathbf{a}^T\mathbf{\Sigma_y}\mathbf{a}} \quad s.t. \quad \mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}}) = 1, \tag{6}$$

and setting b *to the value*

$$b_* = \mathbf{a}_*^T\bar{\mathbf{x}} - \kappa_*\sqrt{\mathbf{a}_*^T\mathbf{\Sigma_x}\mathbf{a}_*},$$

where $\mathbf{a}_*$ *is an optimal solution of (6). The optimal worst-case misclassification probability is obtained via*

$$1 - \alpha_* = \frac{1}{1+\kappa_*^2} = \frac{\left(\sqrt{\mathbf{a}_*^T\mathbf{\Sigma_x}\mathbf{a}_*} + \sqrt{\mathbf{a}_*^T\mathbf{\Sigma_y}\mathbf{a}_*}\right)^2}{1 + \left(\sqrt{\mathbf{a}_*^T\mathbf{\Sigma_x}\mathbf{a}_*} + \sqrt{\mathbf{a}_*^T\mathbf{\Sigma_y}\mathbf{a}_*}\right)^2}.$$

If either $\mathbf{\Sigma_x}$ *or* $\mathbf{\Sigma_y}$ *is positive definite, the optimal hyperplane is unique.*

Once an optimal hyperplane – called a minimax probabilistic decision hyperplane – is found, classification of a new data point $\mathbf{z}_{new}$ is done by evaluating $\operatorname{sign}(\mathbf{a}_*^T \mathbf{z}_{new} - b_*)$: if this is $+1$, $\mathbf{z}_{new}$ is classified as belonging to class $\mathbf{x}$, otherwise $\mathbf{z}_{new}$ is classified as belonging to class $\mathbf{y}$. This algorithm for binary classification is called the Minimax Probability Machine (MPM) for binary classification.

Proof of theorem: Consider the second constraint in (1), with given $\mathbf{a} \neq 0$. Using Lemma 1, this constraint can be expressed as

$$b - \mathbf{a}^T \bar{\mathbf{y}} \geq \kappa(\alpha) \sqrt{\mathbf{a}^T \mathbf{\Sigma}_\mathbf{y} \mathbf{a}} \qquad \text{where} \quad \kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}. \tag{7}$$

We can handle the first constraint in (1) in a similar way (just write $\mathbf{a}^T \mathbf{x} \leq b$ as $-\mathbf{a}^T \mathbf{x} \geq -b$ and apply the result (7)). Notice that we can indeed apply Lemma 1 in both cases: when maximizing α, we prefer a hyperplane for which $\mathbf{a}^T \bar{\mathbf{y}} \leq b \leq \mathbf{a}^T \bar{\mathbf{x}}$, since otherwise $\alpha = 0$.

The optimization problem (1) then becomes:

$$\max_{\alpha, \mathbf{a} \neq 0, b} \ \alpha \quad \text{s.t.} \quad -b + \mathbf{a}^T \bar{\mathbf{x}} \geq \kappa(\alpha) \sqrt{\mathbf{a}^T \mathbf{\Sigma}_\mathbf{x} \mathbf{a}} \tag{8}$$
$$b - \mathbf{a}^T \bar{\mathbf{y}} \geq \kappa(\alpha) \sqrt{\mathbf{a}^T \mathbf{\Sigma}_\mathbf{y} \mathbf{a}}.$$

Since $\kappa(\alpha)$ is a monotone increasing function of α, we can change variables and rewrite our problem as

$$\max_{\kappa, \mathbf{a} \neq 0, b} \ \kappa \quad \text{s.t.} \quad \mathbf{a}^T \bar{\mathbf{y}} + \kappa \sqrt{\mathbf{a}^T \mathbf{\Sigma}_\mathbf{y} \mathbf{a}} \leq b \leq \mathbf{a}^T \bar{\mathbf{x}} - \kappa \sqrt{\mathbf{a}^T \mathbf{\Sigma}_\mathbf{x} \mathbf{a}}, \tag{9}$$

the optimal values of κ and of the worst-case misclassification probability $1 - \alpha$ being related by

$$\alpha_* = \frac{\kappa_*^2}{1 + \kappa_*^2}.$$

Further, we can eliminate b from (9):

$$\max_{\kappa, \mathbf{a} \neq 0} \ \kappa \quad \text{s.t.} \quad \mathbf{a}^T \bar{\mathbf{y}} + \kappa \sqrt{\mathbf{a}^T \mathbf{\Sigma}_\mathbf{y} \mathbf{a}} \leq \mathbf{a}^T \bar{\mathbf{x}} - \kappa \sqrt{\mathbf{a}^T \mathbf{\Sigma}_\mathbf{x} \mathbf{a}}. \tag{10}$$

Since we want to maximize κ, the inequalities in (9) will become equalities at the optimum. An optimal value of b will thus be given by

$$b_* = \mathbf{a}_*^T \bar{\mathbf{x}} - \kappa_* \sqrt{\mathbf{a}_*^T \mathbf{\Sigma}_\mathbf{x} \mathbf{a}_*} = \mathbf{a}_*^T \bar{\mathbf{y}} + \kappa_* \sqrt{\mathbf{a}_*^T \mathbf{\Sigma}_\mathbf{y} \mathbf{a}_*},$$

where $\mathbf{a}_*$ and κ_* are optimal values of $\mathbf{a}$ and κ respectively. Rearranging the constraint in (10), we obtain the optimization problem

$$\max_{\kappa, \mathbf{a} \neq 0} \ \kappa \quad \text{s.t.} \quad \mathbf{a}^T (\bar{\mathbf{x}} - \bar{\mathbf{y}}) \geq \kappa \left(\sqrt{\mathbf{a}^T \mathbf{\Sigma}_\mathbf{x} \mathbf{a}} + \sqrt{\mathbf{a}^T \mathbf{\Sigma}_\mathbf{y} \mathbf{a}} \right). \tag{11}$$

If $\bar{\mathbf{x}} = \bar{\mathbf{y}}$, then $\mathbf{a} \neq 0$ implies $\kappa = 0$, which in turn yields $\alpha = 0$. In this case, the minimax probability decision problem (1) does not have a meaningful solution, and the optimal worst-case misclassification probability is $1 - \alpha_* = 1$.

Let us proceed with the assumption $\bar{\mathbf{x}} \neq \bar{\mathbf{y}}$. We observe that condition (11) is positively homogeneous in $\mathbf{a}$: if $\mathbf{a}$ satisfies (11), $s\mathbf{a}$ with $s \geq 0$ also does. Furthermore, (11) implies $\mathbf{a}^T(\bar{\mathbf{x}}-\bar{\mathbf{y}}) \geq 0$. Since $\bar{\mathbf{x}} \neq \bar{\mathbf{y}}$, we can set $\mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}}) = 1$ without loss of generality. This implies $\mathbf{a} \neq 0$, and in turn, $\sqrt{\mathbf{a}^T\mathbf{\Sigma_x}\mathbf{a}} + \sqrt{\mathbf{a}^T\mathbf{\Sigma_y}\mathbf{a}} \neq 0$. Thus we can write the optimization problem as

$$\max_{\kappa,\mathbf{a}} \ \kappa \quad \text{s.t.} \quad \kappa \leq \left(\sqrt{\mathbf{a}^T\mathbf{\Sigma_x}\mathbf{a}} + \sqrt{\mathbf{a}^T\mathbf{\Sigma_y}\mathbf{a}}\right)^{-1}$$
$$\mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}}) = 1,$$

which allows us to eliminate κ:

$$\min_{\mathbf{a}} \|\mathbf{\Sigma_x}^{1/2}\mathbf{a}\|_2 + \|\mathbf{\Sigma_y}^{1/2}\mathbf{a}\|_2 \quad \text{s.t.} \quad \mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}}) = 1. \tag{12}$$

The above problem is convex, feasible (since $\bar{\mathbf{x}} \neq \bar{\mathbf{y}}$), and its objective is bounded below, therefore there exists an optimal point, $\mathbf{a}_*$. When either $\mathbf{\Sigma_x}$ or $\mathbf{\Sigma_y}$ is positive definite, the strict convexity of the objective function implies that the optimal point is unique. This ends our proof of Theorem 2. □

2.3 Solving the Optimization Problem

Problem (6) is a convex optimization problem, more precisely a second order cone program (SOCP) (Boyd and Vandenberghe, 2001). General-purpose programs such as SeDuMi (Sturm, 1999) or Mosek (Andersen and Andersen, 2000) can handle those problems efficiently. These codes use interior-point methods for SOCP (Nesterov and Nemirovsky, 1994, Lobo et al., 1998), which yield a worst-case complexity of $O(n^3)$. Of course, the first and second moments of $\mathbf{x}, \mathbf{y}$ must be estimated beforehand, using for example sample moment plug-in estimates $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{\Sigma}}_\mathbf{x}, \hat{\mathbf{\Sigma}}_\mathbf{y}$ for $\bar{\mathbf{x}}, \bar{\mathbf{y}}, \mathbf{\Sigma_x}$, and $\mathbf{\Sigma_y}$ respectively. This brings the total complexity to $O(n^3 + Nn^2)$, where N is the number of data points. This is the same complexity as the quadratic programs one has to solve in linear support vector machines (Schölkopf and Smola, 2002).

To gain more insight into the nature of the problem, we propose the following simple and perhaps more transparent iterative least-squares method to globally solve the problem. Similar iterative procedures to solve weighted least-squares problems have been applied before in the SVM literature to achieve an approximate solution as in the Least Squares SVM described by Suykens and Vandewalle (1999), or an exact solution as by Pérez-Cruz et al. (2001).

Notice that we can write $\mathbf{a} = \mathbf{a}_0 + \mathbf{F}\mathbf{u}$, where $\mathbf{u} \in \mathbb{R}^{n-1}$, $\mathbf{a}_0 = (\bar{\mathbf{x}}-\bar{\mathbf{y}})/\|\bar{\mathbf{x}}-\bar{\mathbf{y}}\|_2^2$, and $\mathbf{F} \in \mathbb{R}^{n\times(n-1)}$ is an orthogonal matrix whose columns span the subspace of vectors orthogonal to $\bar{\mathbf{x}} - \bar{\mathbf{y}}$. Hence, we can eliminate the constraint in (12) and write this as an unconstrained SOCP:

$$\min_{\mathbf{u}} \|\mathbf{\Sigma_x}^{1/2}(\mathbf{a}_0 + \mathbf{F}\mathbf{u})\|_2 + \|\mathbf{\Sigma_y}^{1/2}(\mathbf{a}_0 + \mathbf{F}\mathbf{u})\|_2.$$

An equivalent form is

$$\inf_{\mathbf{u},\beta>0,\eta>0} \quad \beta + \frac{1}{\beta}\|\mathbf{\Sigma_x}^{1/2}(\mathbf{a}_0 + \mathbf{F}\mathbf{u})\|_2^2 + \eta + \frac{1}{\eta}\|\mathbf{\Sigma_y}^{1/2}(\mathbf{a}_0 + \mathbf{F}\mathbf{u})\|_2^2. \tag{13}$$

This equivalence can be understood by fixing $\mathbf{u}$ and minimizing the objective in (13) with respect to β and η. Notice that a factor $\frac{1}{2}$ over each term has been dropped in (13). This is irrelevant for the optimal value of $\mathbf{u}$, β and η, although it has to be taken into account when computing κ_* and α_*.

The function to be minimized is jointly convex in $\mathbf{u}$, β, η (as follows easily from the convexity of the function $f(x,t) = x^2/t$ over the domain $\{(x,t) \ : \ x \in \mathbb{R}, \ t > 0\}$). To minimize (13), we used an iterative least-squares approach based on the above formulation, with regularization of the Hessian for computational stability. The algorithm, which is presented in high-level pseudocode in Table 1, is an instance of "Block Coordinate Descent" (Bertsekas, 1999). At iteration k, we first minimize with respect to β, η by setting $\beta_k = \|{\mathbf{\Sigma_x}}^{1/2}(\mathbf{a}_0 + \mathbf{F}\mathbf{u}_{k-1})\|_2$ (β-step) and $\eta_k = \|{\mathbf{\Sigma_y}}^{1/2}(\mathbf{a}_0 + \mathbf{F}\mathbf{u}_{k-1})\|_2$ (η-step). Those minimizers are unique. Then we minimize with respect to $\mathbf{u}$ by solving a least-squares problem in $\mathbf{u}$ with β, η fixed ($\mathbf{u}$-step). As already mentioned, this least-squares step is regularized: a regularization term $\delta\mathbf{I}$ (with $\delta > 0$ small) is added to the Hessian of the least-squares step. This makes the Hessian positive definite. Thus, for the $\mathbf{u}$-step also, the minimum is uniquely attained. As shown by Bertsekas (1999), these features imply the convergence of the above block coordinate descent method.

2.4 Geometric Interpretation

Problem (6) admits an appealing geometric interpretation, which we obtain via convex duality. To address a meaningful problem we assume $\bar{\mathbf{x}} \neq \bar{\mathbf{y}}$ and for simplicity we further assume that either $\mathbf{\Sigma_x}$ or $\mathbf{\Sigma_y}$ is positive definite.

To obtain the dual of the problem, we first express (12) as the following constrained minimax problem:

$$\lambda_* := \min_{\mathbf{a}} \max_{\lambda,\mathbf{u},\mathbf{v}} \mathbf{u}^T {\mathbf{\Sigma_x}}^{1/2}\mathbf{a} + \mathbf{v}^T {\mathbf{\Sigma_y}}^{1/2}\mathbf{a} + \lambda(1 - \mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}})) \ : \ \|\mathbf{u}\|_2 \leq 1, \ \|\mathbf{v}\|_2 \leq 1.$$

The above is based on the following expression for the 2-norm of a vector $\mathbf{z}$: $\|\mathbf{z}\|_2 = \max\{\mathbf{u}^T\mathbf{z} \ : \ \|\mathbf{u}\|_2 \leq 1\}$. Exchanging the min and max operators in the above yields the lower bound

$$\lambda_* \geq \max_{\lambda,\mathbf{u},\mathbf{v}} g(\lambda, \mathbf{u}, \mathbf{v}) \ : \ \|\mathbf{u}\|_2 \leq 1, \ \|\mathbf{v}\|_2 \leq 1,$$

where g is the so-called dual function:

$$g(\mathbf{u}, \mathbf{v}, \lambda) = \max_{\mathbf{a}} \mathbf{u}^T {\mathbf{\Sigma_x}}^{1/2}\mathbf{a} + \mathbf{v}^T {\mathbf{\Sigma_y}}^{1/2}\mathbf{a} + \lambda(1 - \mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}})) = \begin{cases} \lambda & \text{if } \lambda\bar{\mathbf{x}} - {\mathbf{\Sigma_x}}^{1/2}\mathbf{u} = \lambda\bar{\mathbf{y}} + {\mathbf{\Sigma_y}}^{1/2}\mathbf{v} \\ -\infty & \text{otherwise.} \end{cases}$$

We obtain the dual problem as

$$\max_{\lambda,\mathbf{u},\mathbf{v}} \lambda \ : \ \|\mathbf{u}\|_2 \leq 1, \ \ \|\mathbf{v}\|_2 \leq 1, \ \ \lambda\bar{\mathbf{x}} - {\mathbf{\Sigma_x}}^{1/2}\mathbf{u} = \lambda\bar{\mathbf{y}} + {\mathbf{\Sigma_y}}^{1/2}\mathbf{v}.$$

Our assumptions ($\bar{\mathbf{x}} \neq \bar{\mathbf{y}}$ and either $\mathbf{\Sigma_x}$ or $\mathbf{\Sigma_y}$ positive definite) imply that the above is strictly feasible. Since the original primal problem also is, we can invoke standard convex duality theorems (see e.g., Boyd and Vandenberghe, 2001) and conclude that both problems have the same optimal values, and both are attained. Further, positive definiteness of $\mathbf{\Sigma_x}$ (or $\mathbf{\Sigma_y}$) guarantees that the optimal value is non zero. Hence, we may make the change of variable $\kappa := 1/\lambda$ in the dual problem and rewrite it as

$$\min_{\kappa,\mathbf{u},\mathbf{v}} \kappa \ : \ \bar{\mathbf{x}} + {\mathbf{\Sigma_x}}^{1/2}\mathbf{u} = \bar{\mathbf{y}} + {\mathbf{\Sigma_y}}^{1/2}\mathbf{v}, \ \ \|\mathbf{u}\|_2 \leq \kappa, \ \ \|\mathbf{v}\|_2 \leq \kappa. \tag{14}$$

```
Get estimates                  $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{\Sigma}}_{\mathbf{x}}, \hat{\mathbf{\Sigma}}_{\mathbf{y}}$

Compute                        $\mathbf{a}_0 \leftarrow (\hat{\mathbf{x}} - \hat{\mathbf{y}})/\|\hat{\mathbf{x}} - \hat{\mathbf{y}}\|_2^2$
                               $\mathbf{F} \in \mathbb{R}^{n \times (n-1)}$ (an orthogonal matrix whose columns span
                               the subspace of vectors orthogonal to $\hat{\mathbf{x}} - \hat{\mathbf{y}}$)
                               $\mathbf{G} \leftarrow \mathbf{F}^T \hat{\mathbf{\Sigma}}_{\mathbf{x}} \mathbf{F}$
                               $\mathbf{H} \leftarrow \mathbf{F}^T \hat{\mathbf{\Sigma}}_{\mathbf{y}} \mathbf{F}$
                               $\mathbf{g} \leftarrow \mathbf{F}^T \hat{\mathbf{\Sigma}}_{\mathbf{x}} \mathbf{a}_0$
                               $\mathbf{h} \leftarrow \mathbf{F}^T \hat{\mathbf{\Sigma}}_{\mathbf{y}} \mathbf{a}_0$

Initialize                     $\beta_1 = 1,\ \eta_1 = 1,\ k = 1$

Repeat                         $\mathbf{M}_{LS} \leftarrow \frac{1}{\beta_k}\mathbf{G} + \frac{1}{\eta_k}\mathbf{H} + \delta\mathbf{I}$
                               $\mathbf{b}_{LS} \leftarrow -\left(\frac{1}{\beta_k}\mathbf{g} + \frac{1}{\eta_k}\mathbf{h}\right)$
                               solve $\mathbf{M}_{LS}\mathbf{u}_k = \mathbf{b}_{LS}$ w.r.t. $\mathbf{u}_k$
                               $\mathbf{a}_k \leftarrow \mathbf{a}_0 + \mathbf{F} * \mathbf{u}_k$
                               $\beta_{k+1} \leftarrow \sqrt{\mathbf{a}_k^T \hat{\mathbf{\Sigma}}_{\mathbf{x}} \mathbf{a}_k}$
                               $\eta_{k+1} \leftarrow \sqrt{\mathbf{a}_k^T \hat{\mathbf{\Sigma}}_{\mathbf{y}} \mathbf{a}_k}$
                               $k \leftarrow k + 1$
Until stop criterion satisfied

Assign                         $\mathbf{a} \leftarrow \mathbf{a}_{k-1}$
                               $b \leftarrow \mathbf{a}^T \hat{\mathbf{x}} - \frac{\beta_k}{\beta_k + \eta_k}$
                               $\kappa \leftarrow \frac{1}{\beta_k + \eta_k}$
                               $\alpha \leftarrow \frac{\kappa^2}{1+\kappa^2}$
```

Table 1: Algorithmic description of the iterative procedure to solve the optimization problem for a linear MPM. The stop criterion can be chosen by the user, e.g., stopping when the relative change in $\beta_k + \eta_k$ is small enough or a maximum number of iterations has been reached. A Matlab implementation of this algorithm is available on *http://robotics.eecs.berkeley.edu/˜gert/*.

At the optimum, we have

$$\lambda_* = \|{\boldsymbol\Sigma_{\mathbf{x}}}^{1/2}\mathbf{a}_*\|_2 + \|{\boldsymbol\Sigma_{\mathbf{y}}}^{1/2}\mathbf{a}_*\|_2 = 1/\kappa_*.$$

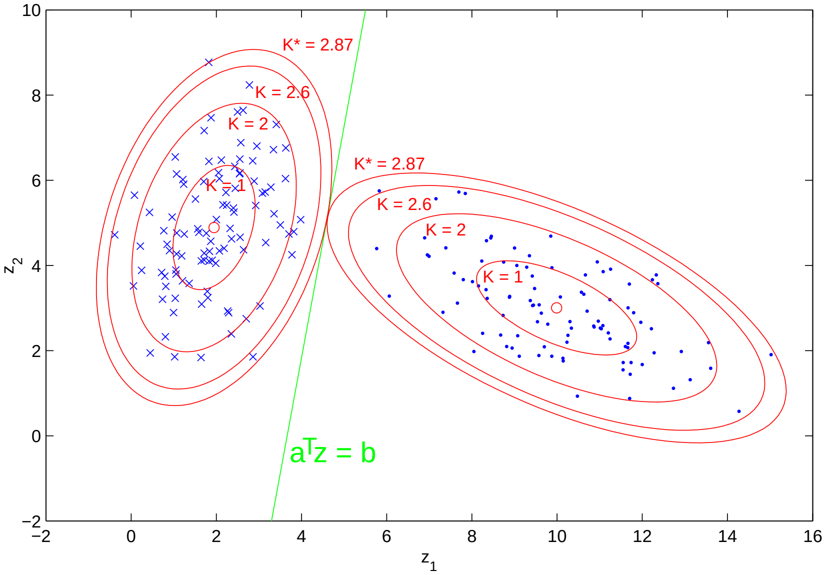


Figure 1: *Geometric interpretation of the minimax probability machine. To find the minimax probabilistic decision hyperplane, consider non-intersecting ellipsoids centered around the means of each class, shape determined by the covariance matrices of each class, which are growing at the same rate, controlled by increasing κ in (15). Finding the optimal decision hyperplane corresponds to finding the smallest $\kappa = \kappa_*$ for which these growing ellipsoids intersect (14). For the optimal κ_*, the ellipsoids will be tangent to each other, and their common tangent is the minimax probabilistic decision hyperplane. The intersection point has the same Mahalanobis distance κ_* to the two classes, and it minimizes the maximum of the Mahalanobis distances to the two classes, as explained in Section 2.5. The optimal worst-case misclassification probability can be read directly from the figure as $1 - \alpha_* = 1/(1 + \kappa_*^2)$.*

In the dual form (14), our problem admits the following geometric interpretation. For given $\kappa \geq 0$, consider two ellipsoids centered around the means of the two classes, and shape determined by their covariance matrices:

$$\mathcal{E}_{\mathbf{x}}(\kappa) = \{\mathbf{x} = \bar{\mathbf{x}} + {\boldsymbol\Sigma_{\mathbf{x}}}^{1/2}\mathbf{u} \ : \ \|\mathbf{u}\|_2 \leq \kappa\}, \ \ \mathcal{E}_{\mathbf{y}}(\kappa) = \{\mathbf{y} = \bar{\mathbf{y}} + {\boldsymbol\Sigma_{\mathbf{y}}}^{1/2}\mathbf{v} \ : \ \|\mathbf{v}\|_2 \leq \kappa\}. \qquad (15)$$

These sets correspond to the points whose Mahalanobis distances to the class means are less than a specified number κ.

Clearly, for κ large enough, these sets intersect. Problem (14) amounts to finding the smallest κ for which these ellipsoids intersect. For the optimal κ, the ellipsoids will be tangent to each other. The minimax probabilistic decision hyperplane is then the common tangent to the optimal ellipsoids. This is illustrated in Figure 1.

2.5 Additional remarks

Another interpretation of our problem assumes a data generation mechanism with bounded uncertainty, where data points in each class are chosen arbitrarily within the ellipsoids $\mathcal{E}_{\mathbf{x}}(\kappa)$ and $\mathcal{E}_{\mathbf{y}}(\kappa)$, defined by (15). We are then asking if there exists a robust linear separator; that is, a hyperplane such that for *any* choice of the data points within their respective ellipsoid, the classifier produces the correct classification. The answer is yes, if the ellipsoids do not intersect, and no otherwise. The minimax probability separator finds the largest κ for which robust separation (in the above sense) is feasible. Thus, we can restate the problem as finding the smallest κ such that $\mathcal{E}_{\mathbf{x}}(\kappa)$ and $\mathcal{E}_{\mathbf{y}}(\kappa)$ intersect:

$$\min_{\kappa,\mathbf{z}} \kappa \ : \ \kappa^2 \geq (\mathbf{z}-\bar{\mathbf{x}})^T {\mathbf{\Sigma}_{\mathbf{x}}}^{-1}(\mathbf{z}-\bar{\mathbf{x}}), \ \ \kappa^2 \geq (\mathbf{z}-\bar{\mathbf{y}})^T {\mathbf{\Sigma}_{\mathbf{y}}}^{-1}(\mathbf{z}-\bar{\mathbf{y}}),$$

or, alternatively, as minimizing the maximum of the Mahalanobis distances to the two classes:

$$\kappa_* = \min_{\mathbf{z}} \max \left(\|{\mathbf{\Sigma}_{\mathbf{x}}}^{-1/2}(\mathbf{z}-\bar{\mathbf{x}})\|_2, \|{\mathbf{\Sigma}_{\mathbf{y}}}^{-1/2}(\mathbf{z}-\bar{\mathbf{y}})\|_2 \right).$$

At optimum, the intersection point has the same Mahalanobis distance to both classes. The optimal worst-case misclassification probability is related to the above optimal value κ_* by $1-\alpha_* = 1/(1+\kappa_*^2)$.

An interesting analogy to this result can be found in SVM classification (Bennett and Bredensteiner, 2000, Crisp and Burges, 1999). In that case, the dual to the problem of finding the maximal margin is the problem of finding points in the convex hulls of the two classes that are closest. This corresponds to a quadratic program (QP) instead of a second order cone program (SOCP).

Another interesting connection with maximal margin classification can be found when viewing the constraints of the original problem (1) in the light of Lemma 1. The constraint (3) is very similar to the constraints in hard margin SVMs. In the MPM case, we have one constraint per class (instead of as many constraints as there are data points in SVMs), where $\kappa(\alpha)$ corresponds to the margin (normalized to one in SVMs) and an extra factor $\sqrt{\mathbf{a}^T\mathbf{\Sigma}\mathbf{a}}$ is added for each class. Thus, minimax probabilistic classification is similar to maximum margin classification with respect to the mean of the classes, where a factor that depends on the covariance matrices of each of the classes pushes the threshold towards the class with lower covariance. Unlike support vector classification, for which the points close to the decision boundary are most important, the MPM looks at the margin between the means of both classes, which rather represent the "typical" examples of each of the classes. By paying more attention to the "typical" rather than the boundary points, the MPM is in some sense similar to the relevance vector machine proposed in Tipping (2000). Furthermore, the MPM is related to vicinal risk minimization (Chapelle et al., 2001, Vapnik, 1999), in which SVMs were "improved" using the covariance of the classes to push the hyperplane away from the samples that belong to the class with the largest covariance matrix.

2.6 Making Gaussian Assumptions

It is interesting to see what happens if we make distributional assumptions; in particular, let us assume that $\mathbf{x} \sim \mathcal{N}(\bar{\mathbf{x}}, \mathbf{\Sigma_x})$ and $\mathbf{y} \sim \mathcal{N}(\bar{\mathbf{y}}, \mathbf{\Sigma_y})$. The first constraint in (1) then becomes:

$$\begin{aligned} \inf_{\mathbf{x}\sim\mathcal{N}(\bar{\mathbf{x}},\mathbf{\Sigma_x})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x} \geq b\} &= \mathbf{Pr}\left\{\mathcal{N}(0,1) \geq \frac{b - \mathbf{a}^T\bar{\mathbf{x}}}{\sqrt{\mathbf{a}^T\mathbf{\Sigma_x}\mathbf{a}}}\right\} \\ &= 1 - \Phi\left(\frac{b - \mathbf{a}^T\bar{\mathbf{x}}}{\sqrt{\mathbf{a}^T\mathbf{\Sigma_x}\mathbf{a}}}\right) \\ &= \Phi\left(\frac{-b + \mathbf{a}^T\bar{\mathbf{x}}}{\sqrt{\mathbf{a}^T\mathbf{\Sigma_x}\mathbf{a}}}\right) \;\geq\; \alpha, \end{aligned}$$

where $\Phi(z)$ is the cumulative distribution function for a standard normal Gaussian distribution:

$$\Phi(z) = \mathbf{Pr}\left\{\mathcal{N}(0,1) \leq z\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} \exp(-s^2/2)\, ds.$$

Since $\Phi(z)$ is monotone increasing, we can write the first constraint in (1) as

$$-b + \mathbf{a}^T\bar{\mathbf{x}} \geq \Phi^{-1}(\alpha)\sqrt{\mathbf{a}^T\mathbf{\Sigma_x}\mathbf{a}}.$$

The same can be done for the second constraint in (1). This leads to the following optimization problem:

$$\begin{aligned} \max_{\alpha,\mathbf{a}\neq 0,b} \;\alpha \quad \text{s.t.} \quad & -b + \mathbf{a}^T\bar{\mathbf{x}} \geq \Phi^{-1}(\alpha)\sqrt{\mathbf{a}^T\mathbf{\Sigma_x}\mathbf{a}} \\ & b - \mathbf{a}^T\bar{\mathbf{y}} \geq \Phi^{-1}(\alpha)\sqrt{\mathbf{a}^T\mathbf{\Sigma_y}\mathbf{a}}. \end{aligned}$$

This has the same form as (8), but now with $\kappa(\alpha) = \Phi^{-1}(\alpha)$ instead of $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$. α again disappears from the optimization problem because $\kappa(\alpha)$ is monotone increasing. We thus solve the *same* optimization problem and find the same decision hyperplane $\mathbf{a}^T\mathbf{z} = b$. The only difference lies in the *value* of the worst-case misclassification probability $1 - \alpha$ associated with κ_*: α will be higher under Gaussian assumptions ($\alpha^*_{\text{gauss}} = \Phi(\kappa_*)$), so the hyperplane will have a higher predicted probability of classifying future data correctly. This is expected: the extra knowledge resulting from the Gaussian assumption should indeed allow us to predict a higher probability of correct future classification.

2.7 Link with Fisher Discriminant Analysis

A discriminant hyperplane based on the first two moments can also be computed via Fisher discriminant analysis (FDA). This involves solving the following optimization problem (Fukunaga, 1990):

$$\max_{\mathbf{a}} \quad \kappa_{\text{FDA}}(\mathbf{a}) := \frac{|\mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}})|}{\sqrt{\mathbf{a}^T\mathbf{\Sigma_x}\mathbf{a} + \mathbf{a}^T\mathbf{\Sigma_y}\mathbf{a}}}. \tag{16}$$

The optimal $\mathbf{a}$ for the FDA cost function, which we denote as $\mathbf{a}_*^{\text{FDA}}$, corresponds to a direction which gives good separation between the two projected sets $\mathbf{a}^T\mathbf{x}$ and $\mathbf{a}^T\mathbf{y}$ with small projected variances. However it is not known whether this optimal direction $\mathbf{a}_*^{\text{FDA}}$ can be used to compute a bound on the generalization error.

On the other hand, the minimax hyperplane is obtained by solving the following optimization problem:

$$\max_{\mathbf{a}} \quad \kappa_{\text{MPM}}(\mathbf{a}) := \frac{|\mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}})|}{\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{a}} + \sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{a}}}.$$

As we have seen, this optimization problem is motivated theoretically as the minimization of a worst-case misclassification probability. Optimizing the MPM cost function also involves seeking an optimal direction $\mathbf{a}$, denoted $\mathbf{a}_*^{\text{MPM}}$, in which we have good separation between the two projected sets $\mathbf{a}^T\mathbf{x}$ and $\mathbf{a}^T\mathbf{y}$ with small projected variances, but the way in which the means and covariances are combined is different in MPM and FDA .

¿From the proof of Theorem 2, we also know that for a given $\mathbf{a}$ such that

$$\kappa \leq \frac{\mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}})}{\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{a}} + \sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{a}}},$$

the worst-case misclassification probability $\beta(\mathbf{a})$ associated with a separating hyperplane $\mathcal{H}(\mathbf{a}, b)$ is bounded above by $1 - \alpha = \frac{1}{1+\kappa^2}$, if b is feasible for problem (8), that is:

$$\mathbf{a}^T\bar{\mathbf{y}} + \kappa\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{a}} \leq b \leq \mathbf{a}^T\bar{\mathbf{x}} - \kappa\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{a}}.$$

To obtain the smallest possible upper bound we let κ increase to $\kappa_{\text{MPM}}(\mathbf{a})$ for a given $\mathbf{a}$. This implies that for every $\mathbf{a} \neq 0$, the misclassification probability for the hyperplane $\mathcal{H}(\mathbf{a}, b)$ is bounded above as follows:

$$\beta(\mathbf{a}) \leq \frac{1}{1 + \kappa_{\text{MPM}}(\mathbf{a})^2},$$

provided $b = \mathbf{a}^T\bar{\mathbf{x}} - \kappa_{\text{MPM}}(\mathbf{a})\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{a}}$.

Using the fact that $\sqrt{2}\kappa_{\text{MPM}}(\mathbf{a}) \geq \kappa_{\text{FDA}}(\mathbf{a})$, we obtain an upper bound on the generalization error for FDA:

$$\beta(\mathbf{a}_*^{\text{FDA}}) \leq \frac{1}{1 + \kappa_{\text{MPM}}(\mathbf{a}_*^{\text{FDA}})^2} \leq \frac{1}{1 + 0.5\kappa_{\text{FDA}}(\mathbf{a}_*^{\text{FDA}})^2}, \tag{17}$$

provided we discriminate between points using a decision hyperplane $\mathcal{H}(\mathbf{a}_*^{\text{FDA}}, b_*^{\text{FDA}})$ where

$$b_*^{\text{FDA}} = (\mathbf{a}_*^{\text{FDA}})^T\bar{\mathbf{x}} - \kappa_{\text{MPM}}(\mathbf{a}_*^{\text{FDA}})\sqrt{(\mathbf{a}_*^{\text{FDA}})^T\boldsymbol{\Sigma}_{\mathbf{x}}(\mathbf{a}_*^{\text{FDA}})}.$$

This not only gives us a bound on the FDA generalization error, it also gives us a way to optimally determine the intercept b for FDA, without assuming Gaussian class-conditional distributions. In case of equal covariance, $\boldsymbol{\Sigma}_{\mathbf{x}} = \boldsymbol{\Sigma}_{\mathbf{y}}$, the two bounds in (17) are equal, and FDA and MPM provide identical classifiers.

Fisher discriminant analysis is essentially a feature extraction method that aids in classification without specifically aiming at solving the classification problem. The FDA criterion function κ_{FDA} is built upon the intuition that "separation" (as characterized by first and second moments) is useful, and upon the desideratum of computational efficiency. Indeed, the solution of (16) is found by solving a generalized eigenvalue problem. The criterion function for MPM, on the other hand, differs from the FDA criterion function in detail, but due to its similar form preserves the intuition of finding a direction that separates the data. It does so while aiming directly at solving the classification problem, and yields an algorithm with similar complexity.

3 Robustness to Estimation Errors

In practical experiments, it may well happen that the error rate computed on the test set is greater than $1 - \alpha_*$. This seems to contradict the previous claim that $1 - \alpha_*$ is an upper bound on misclassification error.

This apparent paradox has to do with estimation errors. Since we do not know the mean and covariance *a priori*, they need to be estimated from the data. The validity of the bound depends on how good the estimate is. This is an issue especially in the "small sample" case. In this section we outline an approach to the estimation problem based on robust optimization.

3.1 The Robust Decision Problem

We start from the original problem as defined in (1) and modify it as follows. We now assume that the mean and covariance matrix of each class are only known within some specified set. In particular, we assume that $(\bar{\mathbf{x}}, \mathbf{\Sigma_x}) \in \mathcal{X}$, where $\mathcal{X}$ is a subset of $\mathbb{R}^n \times \mathcal{S}_n^+$, where $\mathcal{S}_n^+$ is the set of $n \times n$ symmetric, positive semidefinite matrices. Likewise we define a set $\mathcal{Y}$ describing uncertainty in the mean and covariance matrix of the random variable $\mathbf{y}$.

We pose the robust counterpart to the original problem, defined as:

$$\begin{aligned} \max_{\alpha, \mathbf{a} \neq 0, b} \ \alpha \quad \text{s.t.} \quad & \inf_{\mathbf{x} \sim (\bar{\mathbf{x}}, \mathbf{\Sigma_x})} \mathbf{Pr}\{\mathbf{a}^T \mathbf{x} \geq b\} \geq \alpha \ \ \forall\, (\bar{\mathbf{x}}, \mathbf{\Sigma_x}) \in \mathcal{X}, \\ & \inf_{\mathbf{x} \sim (\bar{\mathbf{y}}, \mathbf{\Sigma_y})} \mathbf{Pr}\{\mathbf{a}^T \mathbf{y} \leq b\} \geq \alpha \ \ \forall\, (\bar{\mathbf{y}}, \mathbf{\Sigma_y}) \in \mathcal{Y}. \end{aligned} \tag{18}$$

In other words, we would like to guarantee a worst-case misclassification probability for all distributions which have unknown-but-bounded mean and covariance matrix, but are otherwise arbitrary.

Using the previous approach it is straightforward to obtain an equivalent formulation of the robust problem:

$$\begin{aligned} \max_{\kappa, \mathbf{a} \neq 0, b} \ \kappa \quad \text{s.t.} \quad & \forall\, (\bar{\mathbf{x}}, \mathbf{\Sigma_x}) \in \mathcal{X}, \ \ -b + \mathbf{a}^T \bar{\mathbf{x}} \geq \kappa \sqrt{\mathbf{a}^T \mathbf{\Sigma_x} \mathbf{a}} \\ & \forall\, (\bar{\mathbf{y}}, \mathbf{\Sigma_y}) \in \mathcal{Y}, \ \ b - \mathbf{a}^T \bar{\mathbf{y}} \geq \kappa \sqrt{\mathbf{a}^T \mathbf{\Sigma_y} \mathbf{a}}. \end{aligned} \tag{19}$$

3.2 A Specific Uncertainty Model

The complexity of the above problem depends of course on the structure of the sets $\mathcal{X}$ and $\mathcal{Y}$. We now consider a specific choice for these sets, that is both realistic from a statistical viewpoint and tractable numerically. Specifically, we consider

$$\begin{aligned} \mathcal{X} &= \left\{ (\bar{\mathbf{x}}, \mathbf{\Sigma_x}) \ : \ (\bar{\mathbf{x}} - \bar{\mathbf{x}}^0)^T {\mathbf{\Sigma_x}}^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{x}}^0) \leq \nu^2, \ \ \|\mathbf{\Sigma_x} - {\mathbf{\Sigma_x}}^0\|_F \leq \rho \right\}, \\ \mathcal{Y} &= \left\{ (\bar{\mathbf{y}}, \mathbf{\Sigma_y}) \ : \ (\bar{\mathbf{y}} - \bar{\mathbf{y}}^0)^T {\mathbf{\Sigma_y}}^{-1} (\bar{\mathbf{y}} - \bar{\mathbf{y}}^0) \leq \nu^2, \ \ \|\mathbf{\Sigma_y} - {\mathbf{\Sigma_y}}^0\|_F \leq \rho \right\}. \end{aligned} \tag{20}$$

Here, $\nu \geq 0$ and $\rho \geq 0$ are fixed (to simplify, we assume them to be the same for both $\mathcal{X}$ and $\mathcal{Y}$). The notation $\bar{\mathbf{x}}^0, \mathbf{\Sigma_x}^0$ stands for our "nominal" estimate of the mean and covariance matrix, respectively. The matrix norm involved in the above is the Frobenius norm: $\|A\|_F^2 = \mathbf{Tr}(A^T A)$.

In our model, the mean of class $\mathbf{x}$ belongs to an ellipsoid centered around $\bar{\mathbf{x}}^0$, with shape determined by the (unknown) $\mathbf{\Sigma_x}$. The covariance matrix itself belongs to a matrix norm ball (in the space of symmetric matrices of order n) centered around $\mathbf{\Sigma_x}^0$. Note that we are not including positive semidefiniteness constraints on the covariance matrices in our model. This means that we are taking a more conservative view and enlarging the uncertainty set. For small values of the uncertainty size ρ, and provided both covariance estimates $\mathbf{\Sigma_x}^0$ and $\mathbf{\Sigma_y}^0$ are positive definite, this assumption is done without harm, since the matrix norm ball will be included in the cone of positive semidefinite matrices.

Our model for the mean uncertainty is motivated by the standard statistical approach to estimating a region of confidence based on Laplace (that is, second-order) approximations to a likelihood function (see e.g., Kass et al., 1988). The model for uncertainty in the covariance matrix is perhaps less classical from a statistical viewpoint, but it leads to a regularization term of a classical form. The specific values of ν and ρ in our model can be determined based on the central limit theorem or on resampling methods.

3.3 Estimation Errors in the Means

We first consider the case where the covariance matrix for each class is known. Later, we will extend our results to the case $\rho > 0$.

For a given $\mathbf{a}$ and $\bar{\mathbf{x}}^0$, we have (see Appendix B)

$$\min_{\bar{\mathbf{x}} \,:\, (\bar{\mathbf{x}}-\bar{\mathbf{x}}^0)^T \mathbf{\Sigma_x}^{-1} (\bar{\mathbf{x}}-\bar{\mathbf{x}}^0) \leq \nu^2} \mathbf{a}^T \bar{\mathbf{x}} = \mathbf{a}^T \bar{\mathbf{x}}^0 - \nu \sqrt{\mathbf{a}^T \mathbf{\Sigma_x} \mathbf{a}}. \tag{21}$$

Using this result, we obtain the following expression for problem (19):

$$\begin{aligned} \max_{\kappa, \mathbf{a} \neq 0, b} \ \kappa \quad \text{s.t.} \quad & -b + \mathbf{a}^T \bar{\mathbf{x}}^0 \geq (\kappa + \nu) \sqrt{\mathbf{a}^T \mathbf{\Sigma_x} \mathbf{a}} \\ & b - \mathbf{a}^T \bar{\mathbf{y}}^0 \geq (\kappa + \nu) \sqrt{\mathbf{a}^T \mathbf{\Sigma_y} \mathbf{a}}. \end{aligned} \tag{22}$$

Note that the optimization problem (8) and (22) are actually the same; only the optimal values of κ differ. Let κ_*^{-1} be the optimal value of problem (6). Then the optimal solution to the robust version above is given by

$$\kappa_*^{\text{rob}} = \kappa_* - \nu.$$

If the optimal value of the original (non-robust) problem (6) is below ν, that is $\kappa_* < \nu$, we conclude that the robust version is not feasible; in other words, there is no way to find a hyperplane which separates the two classes in the robust minimax probabilistic sense. The worst-case misclassification probability is then $1 - \alpha_*^{\text{rob}} = 1$. If $\kappa_* \geq \nu$, the robust hyperplane is the *same* as the non-robust one; the only change is in the increase in the worst-case misclassification probability, which now is

$$1 - \alpha_*^{\text{rob}} = \frac{1}{1 + (\kappa_* - \nu)^2}.$$

This gives us a new interpretation of the optimal κ_* in problem (6) of Theorem 2. The optimal κ_* is the worst-case uncertainty in the means (in the sense of our ellipsoidal model) that can be tolerated. If the level of uncertainty is more than κ_* then the robust problem becomes infeasible.

This can be understood deterministically via the interpretation given in Section 2.5. Consider again the ellipsoids given by (15). Now assume that the means (that is, the centers of the ellipsoids) are uncertain; in our model, the center $\bar{\mathbf{x}}$ is allowed to vary in the ellipsoid $\mathcal{E}_{\mathbf{x}}(\nu)$, whose shape and orientation is the same as $\mathcal{E}_{\mathbf{x}}(\kappa)$. This simply means that we can reformulate the robust problem as a variant of the original one, where we enlarge $\mathcal{E}_{\mathbf{x}}(\kappa)$ by changing κ to $\kappa+\nu$. This is illustrated in Figure 2.

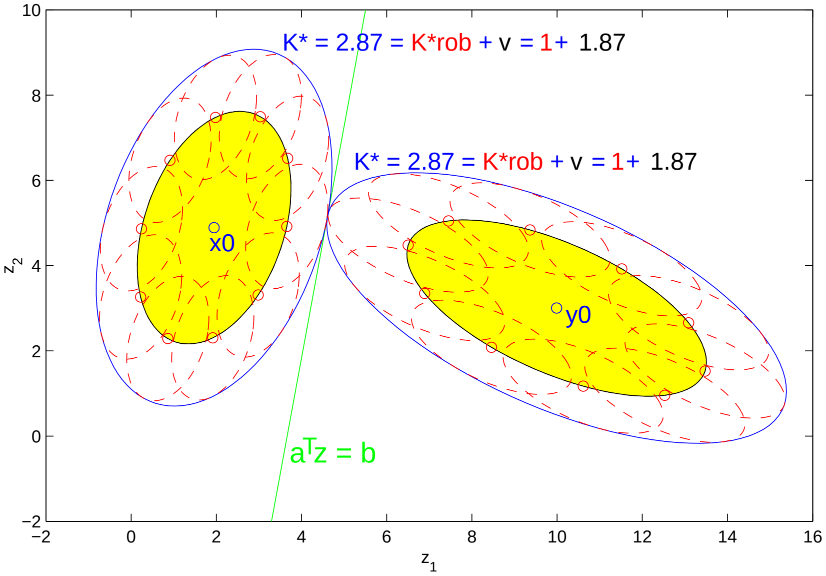


Figure 2: *Geometric interpretation of the result with estimation errors in the means. Again, we consider the ellipsoids $\mathcal{E}_{\mathbf{x}}(\kappa)$ and $\mathcal{E}_{\mathbf{y}}(\kappa)$ given by (15). But now their centers $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are allowed to vary within ellipsoids $\mathcal{E}_{\mathbf{x}}(\nu)$ and $\mathcal{E}_{\mathbf{y}}(\nu)$, with centers $\bar{\mathbf{x}}^0$ and $\bar{\mathbf{y}}^0$, and shape and orientation similar to that of $\mathcal{E}_{\mathbf{x}}(\kappa)$ and $\mathcal{E}_{\mathbf{y}}(\kappa)$, respectively. Because of the similarity in shape and orientation of the κ-ellipsoids and the ν-ellipsoids, we can simply reformulate the robust problem as the original one, where the growth of $\mathcal{E}_{\mathbf{x}}$ and $\mathcal{E}_{\mathbf{y}}$ is now controlled by $\kappa+\nu$ (where ν is fixed and κ varied) instead of κ. The optimal value κ_* for the original problem (6) is linked with the optimal solution κ_*^{rob} to the robust version (22): $\kappa_* = \kappa_*^{\mathrm{rob}} + \nu$.*

3.4 Estimation Errors in the Covariance Matrix

In this section we consider the case where there is uncertainty in the covariance but the mean is accurately estimated from the samples ($\nu = 0$).

To address the robustness conditions in (19), we are led to the following problem:

$$\max_{\mathbf{\Sigma}} \ \mathbf{a}^T\mathbf{\Sigma}\mathbf{a} \ : \ \|\mathbf{\Sigma}-\mathbf{\Sigma}^0\|_F \le \rho,$$

where $\mathbf{a}$ is given for now. The optimal value of the above problem is (see Appendix C):

$$\max_{\mathbf{\Sigma}\,:\,\|\mathbf{\Sigma}-\mathbf{\Sigma}^0\|_F\le\rho} \mathbf{a}^T\mathbf{\Sigma}\mathbf{a} = \mathbf{a}^T\left(\mathbf{\Sigma}^0+\rho I_n\right)\mathbf{a}, \tag{23}$$

where I_n is the $n \times n$ identity matrix.

Problem (19) reads now

$$\max_{\kappa, \mathbf{a} \neq 0, b} \kappa \quad \text{s.t.} \quad -b + \mathbf{a}^T \bar{\mathbf{x}} \geq \kappa \sqrt{\mathbf{a}^T (\mathbf{\Sigma_x} + \rho I_n) \mathbf{a}} \qquad (24)$$
$$b - \mathbf{a}^T \bar{\mathbf{y}} \geq \kappa \sqrt{\mathbf{a}^T (\mathbf{\Sigma_y} + \rho I_n) \mathbf{a}}.$$

The robust problem is thus exactly the same as the original one, with a term ρI_n added to the covariance matrices.

We can interpret the term ρI_n as a regularization term. In the original statement of Theorem 2, if the estimates of the covariance matrices are not positive definite, the original problem is not strictly convex. This leads to potentially many solutions, possible discontinuities, and numerical problems. Adding such a term will make the solution unique and also a continuous function of the input data.

3.5 Summary of Robustness Results

When both $\nu > 0$ and $\rho > 0$, that is, in the case of combined uncertainty for both means and covariance matrices, we simply apply the results pertaining to the uncertainty on the covariance matrices to the modified problem (22). Starting from (22), we obtain a formulation similar to (24), with $\kappa + \nu$ in the constraints instead of κ. Our findings can be summarized as follows.

Theorem 3 *The optimal robust minimax probability classifier from problem (18) with sets $\mathcal{X}$, $\mathcal{Y}$ given by (20) can be obtained by solving problem (6), with $\mathbf{\Sigma_x} = \mathbf{\Sigma_x}^0 + \rho I_n$, $\mathbf{\Sigma_y} = \mathbf{\Sigma_y}^0 + \rho I_n$. If κ_*^{-1} is the optimal value of that problem, the corresponding worst-case misclassification probability is*

$$1 - \alpha_*^{\text{rob}} = \frac{1}{1 + \max(0, (\kappa_* - \nu))^2}.$$

4 Single Class Case

In this section we extend our minimax approach to classification to the single class case, with linear decision boundaries.

The problem addressed in Section 2 is one of supervised learning: for each data point, a label $+1$ or -1 is given, and the goal is to classify future data as belonging to one of these two classes. In the most general case, unsupervised learning essentially involves estimation of the density from which given data points $\mathbf{x}$ are drawn. A simplified version of this problem estimates quantiles of this distribution: for $\alpha \in (0, 1]$, one computes a region $\mathcal{Q}$ such that $\mathbf{Pr}\{\mathbf{x} \in \mathcal{Q}\} = \alpha$ (Schölkopf and Smola, 2002). If α is chosen close to 1, this amounts to outlier detection: most of the data will be contained inside the region $\mathcal{Q}$, and a data point outside $\mathcal{Q}$ can be considered as an outlier. Let us consider data $\mathbf{x} \sim (\bar{\mathbf{x}}, \mathbf{\Sigma_x})$ and the linear case where $\mathcal{Q}$ is a half-space not containing the origin. Our basic question is: is the origin an outlier with respect to the data?

For simplicity, we assume that the covariance matrix $\mathbf{\Sigma_x}$ is positive definite. In practice, this will always be the case, since we will add a regularization term to the covariance matrix. This is discussed at the end of this section. Given $\alpha \in (0, 1)$, we seek a half-space $\mathcal{Q}(\mathbf{a}, b) = \{\mathbf{z} \mid \mathbf{a}^T \mathbf{z} \geq b\}$, with $\mathbf{a} \in \mathbb{R}^n \backslash \{0\}$ and $b \in \mathbb{R}$, and not containing $\mathbf{0}$, such that with probability at least α, the data lies in $\mathcal{Q}$, for every distribution having mean $\bar{\mathbf{x}}$ and covariance matrix $\mathbf{\Sigma_x}$:

$$\inf_{\mathbf{x} \sim (\bar{\mathbf{x}}, \mathbf{\Sigma_x})} \mathbf{Pr}\{\mathbf{a}^T \mathbf{x} \geq b\} \geq \alpha.$$

We want the region $\mathcal{Q}$ to be tight, so we maximize its Mahalanobis distance to the origin:

$$\max_{\mathbf{a} \neq 0, b} \frac{b}{\sqrt{\mathbf{a}^T \mathbf{\Sigma_x} \mathbf{a}}} \quad \text{s.t.} \quad \inf_{\mathbf{x} \sim (\bar{\mathbf{x}}, \mathbf{\Sigma_x})} \mathbf{Pr}\{\mathbf{a}^T \mathbf{x} \geq b\} \geq \alpha. \tag{25}$$

Future points $\mathbf{z}$ for which $\mathbf{a}_*^T \mathbf{z} \leq b_*$ can then be considered as outliers with respect to the region $\mathcal{Q}$, with the worst-case probability of occurrence outside $\mathcal{Q}$ given by $1 - \alpha$.

Next we prove the following result.

Theorem 4 *If we assume that the covariance matrix $\mathbf{\Sigma_x}$ is positive definite, the single-class problem (25) is strictly feasible if and only if $\zeta > \kappa(\alpha)$, where*

$$\zeta := \sqrt{\bar{\mathbf{x}}^T {\mathbf{\Sigma_x}}^{-1} \bar{\mathbf{x}}}, \quad \kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}},$$

and $1 - \alpha$ the worst-case probability of occurrence outside region $\mathcal{Q}$. In this case, the optimal half-space $\mathcal{Q}(\mathbf{a}_*, b_*)$ is unique, and determined by

$$\mathbf{a}_* = \frac{{\mathbf{\Sigma_x}}^{-1} \bar{\mathbf{x}}}{\zeta^2 - \kappa(\alpha)\zeta}, \quad b_* = 1. \tag{26}$$

Proof of theorem: Note that $\mathcal{Q}(\mathbf{a}, b)$ does not contain $\mathbf{0}$ if and only if $b > 0$. Also, the optimization problem (25) is positively homogeneous in $(\mathbf{a}, b)$. Thus, without loss of generality, we can set $b = 1$ in problem (25). Finally, we observe that the change of notation $\mathbf{a} \rightarrow {\mathbf{\Sigma_x}}^{-1/2} \mathbf{a}$, $\bar{\mathbf{x}} \rightarrow {\mathbf{\Sigma_x}}^{1/2} \bar{\mathbf{x}}$ allows us to assume $\mathbf{\Sigma_x} = I$ in the derivations that follow.

Consider the constraint in (25), with $\mathbf{a} \neq 0$. Using the proof of Lemma 1, allows us to obtain an equivalent constraint:

$$\frac{\max((\mathbf{a}^T \bar{\mathbf{x}} - 1), 0)^2}{\mathbf{a}^T \mathbf{a}} \geq \frac{\alpha}{1-\alpha},$$

which, in view of $\alpha \in (0, 1)$, is equivalent to

$$\mathbf{a}^T \bar{\mathbf{x}} - 1 \geq \kappa(\alpha) \|\mathbf{a}\|_2.$$

The above also implies $\mathbf{a} \neq 0$. Hence, the optimization problem (25) can be equivalently written as:

$$\min_{\mathbf{a}} \|\mathbf{a}\|_2 \ : \ \mathbf{a}^T \bar{\mathbf{x}} \geq 1 + \kappa(\alpha) \|\mathbf{a}\|_2. \tag{27}$$

Decomposing $\mathbf{a}$ as $\lambda \bar{\mathbf{x}} + \mathbf{z}$, where the variable $\mathbf{z}$ satisfies $\mathbf{z}^T \bar{\mathbf{x}} = 0$, we easily obtain that at the optimum, $\mathbf{z} = 0$. In other words, the optimal $\mathbf{a}$ is parallel to $\bar{\mathbf{x}}$, in the form $\mathbf{a} = \lambda \bar{\mathbf{x}}$, and the problem reduces to the one-dimensional problem:

$$\min_{\lambda} |\lambda| \|\bar{\mathbf{x}}\|_2 \ : \ \lambda \bar{\mathbf{x}}^T \bar{\mathbf{x}} \geq 1 + \kappa(\alpha) \|\bar{\mathbf{x}}\|_2 |\lambda|.$$

The constraint in the above implies that $\lambda \geq 0$, hence the problem reduces to

$$\min_{\lambda} \lambda \ : \ \lambda(\bar{\mathbf{x}}^T\bar{\mathbf{x}} - \kappa(\alpha)\|\bar{\mathbf{x}}\|_2) \geq 1, \ \ \lambda \geq 0.$$

The results of the theorem follow directly. □

One could consider this single-class problem as a particular case of the two-class problem, where the mean and the covariance of the origin are known, i.e., $\bar{\mathbf{y}} = \mathbf{0}$ and $\mathbf{\Sigma_y} = \mathbf{O}$ – where $\mathbf{0}$ and $\mathbf{O}$ denote respectively the null vector and the null matrix of dimension n. Simplifying and solving (12) in this case will indeed lead to the right orientation of the half-space, i.e., the right value for the optimal $\mathbf{a}_*$. However, as can directly be seen from the geometrical interpretation in Figure 1, it will lead to an offset $b_* = 0$, which is wrong in general.

Remark. The results from Section 3 can readily be extended to the single class case. Let the uncertainty on the mean and the covariance matrix of $\mathbf{x}$ again be given by (20). The worst-case probability of occurrence outside region $\mathcal{Q}$ is still given by $1 - \alpha$. The optimal half-space is unique (indeed, $\mathbf{\Sigma_x} + \rho I_n$ is positive definite) and determined by

$$\mathbf{a}_* = \frac{(\mathbf{\Sigma_x} + \rho I_n)^{-1}\bar{\mathbf{x}}}{\zeta^2 - (\kappa(\alpha) + \nu)\zeta} \qquad \text{with} \quad \zeta = \sqrt{\bar{\mathbf{x}}^T(\mathbf{\Sigma_x} + \rho I_n)^{-1}\bar{\mathbf{x}}},$$

if the choice of α is such that $\kappa(\alpha) + \nu \leq \zeta$ or $\alpha \leq \frac{(\zeta-\nu)^2}{1+(\zeta-\nu)^2}$. Notice that the uncertainty in the covariance matrix $\mathbf{\Sigma_x}$ leads to the typical, well-known regularization for inverting this matrix. If the choice of α is not feasible or if $\bar{\mathbf{x}} = \mathbf{0}$, the problem does not have a solution.

5 Kernelization

In this section we describe the "kernelization" of the minimax approach described in Section 2. We seek to map the problem to a higher-dimensional feature space $\mathbb{R}^f$ via a mapping $\varphi : \mathbb{R}^n \mapsto \mathbb{R}^f$, such that a linear decision boundary $\mathcal{H}(\mathbf{a}, b) = \{\varphi(\mathbf{z}) \in \mathbb{R}^f \mid \mathbf{a}^T\varphi(\mathbf{z}) = b\}$ in the feature space $\mathbb{R}^f$ corresponds to a nonlinear decision boundary $\mathcal{D}(\mathbf{a}, b) = \{\mathbf{z} \in \mathbb{R}^n \mid \mathbf{a}^T\varphi(\mathbf{z}) = b\}$ in the original space $\mathbb{R}^n$ ($\mathbf{a} \in \mathbb{R}^f\backslash\{0\}$ and $b \in \mathbb{R}$).

Let the data be mapped as

$$\begin{aligned} \mathbf{x} \mapsto \varphi(\mathbf{x}) &\sim (\overline{\varphi(\mathbf{x})}, \mathbf{\Sigma}_{\varphi(\mathbf{x})}), \\ \mathbf{y} \mapsto \varphi(\mathbf{y}) &\sim (\overline{\varphi(\mathbf{y})}, \mathbf{\Sigma}_{\varphi(\mathbf{y})}). \end{aligned}$$

A nonlinear decision boundary in $\mathbb{R}^n$ can then be obtained by solving the minimax probability decision problem (1) in the higher-dimensional feature space $\mathbb{R}^f$:

$$\begin{aligned} \max_{\alpha, \mathbf{a} \neq 0, b} \ \alpha \quad \text{s.t.} \quad & \inf_{\varphi(\mathbf{x}) \sim (\overline{\varphi(\mathbf{x})}, \mathbf{\Sigma}_{\varphi(\mathbf{x})})} \mathbf{Pr}\{\mathbf{a}^T\varphi(\mathbf{x}) \geq b\} \geq \alpha \\ & \inf_{\varphi(\mathbf{y}) \sim (\overline{\varphi(\mathbf{y})}, \mathbf{\Sigma}_{\varphi(\mathbf{y})})} \mathbf{Pr}\{\mathbf{a}^T\varphi(\mathbf{y}) \leq b\} \geq \alpha. \end{aligned} \tag{28}$$

To carry out this program, we need to reformulate the minimax problem in terms of a given kernel function $K(\mathbf{z}_1, \mathbf{z}_2) = \varphi(\mathbf{z}_1)^T\varphi(\mathbf{z}_2)$ satisfying Mercer's condition.

5.1 Why the Kernel Trick Works

The kernel trick will only work if problem (28) can be entirely expressed in terms of inner products of the mapped data $\varphi(\mathbf{z})$ only. Fortunately, this is indeed the case (even if we take uncertainty in the means and covariance matrices into account), provided we use an appropriate class of estimates for the means and covariance matrices, as defined in the following corollary to Theorem 2.

Corollary 5 *Let $\{\mathbf{x}_i\}_{i=1}^{N_x}$ and $\{\mathbf{y}_i\}_{i=1}^{N_y}$ be the training data points in the classes corresponding to $\mathbf{x}$ and $\mathbf{y}$ respectively. If $\bar{\mathbf{x}}, \bar{\mathbf{y}}, \mathbf{\Sigma_x}, \mathbf{\Sigma_y}$ can be written as*

$$\begin{aligned}
\bar{\mathbf{x}} &= \sum_{i=1}^{N_x} \lambda_i \mathbf{x}_i, \quad \bar{\mathbf{y}} = \sum_{i=1}^{N_y} \omega_i \mathbf{y}_i, \\
\mathbf{\Sigma_x} &= \rho_x I_n + \sum_{i=1}^{N_x} \Lambda_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T, \\
\mathbf{\Sigma_y} &= \rho_y I_n + \sum_{i=1}^{N_y} \Omega_i (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T,
\end{aligned}$$

where I_n is the identity matrix of dimension n, then the optimal $\mathbf{a}$ will lie in the span of the data points $\{\mathbf{x}_i\}_{i=1}^{N_x}$ and $\{\mathbf{y}_i\}_{i=1}^{N_y}$.

The additional regularization terms $\rho_x I_n$ and $\rho_y I_n$ are useful in light of the robustness results in Section 3.4.

Proof of corollary: We can write any $\mathbf{a}$ as $\mathbf{a} = \mathbf{a}_d + \mathbf{a}_p$, where $\mathbf{a}_d$ is the projection of $\mathbf{a}$ in the span of the data (vector space spanned by all the data points $\{\mathbf{x}_i\}_{i=1}^{N_x}$ and $\{\mathbf{y}_i\}_{i=1}^{N_y}$), whereas $\mathbf{a}_p$ is the orthogonal component to the data. One can then easily check that

$$\begin{aligned}
\sqrt{\mathbf{a}^T \mathbf{\Sigma_x} \mathbf{a}} &= \sqrt{\mathbf{a}_d^T \sum_{i=1}^{N_x} \Lambda_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{a}_d + \rho_x (\mathbf{a}_d^T \mathbf{a}_d + \mathbf{a}_p^T \mathbf{a}_p)}, \\
\sqrt{\mathbf{a}^T \mathbf{\Sigma_y} \mathbf{a}} &= \sqrt{\mathbf{a}_d^T \sum_{i=1}^{N_y} \Omega_i (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T \mathbf{a}_d + \rho_y (\mathbf{a}_d^T \mathbf{a}_d + \mathbf{a}_p^T \mathbf{a}_p)}, \\
\mathbf{a}^T (\bar{\mathbf{x}} - \bar{\mathbf{y}}) &= \mathbf{a}_d^T (\bar{\mathbf{x}} - \bar{\mathbf{y}}),
\end{aligned}$$

because $\mathbf{a}_p^T \mathbf{x}_i = 0, i = 1, \ldots, N_x$, $\mathbf{a}_p^T \mathbf{y}_i = 0, i = 1, \ldots, N_y$ and $\mathbf{a}_d^T \mathbf{a}_p = 0$. So, an orthogonal component $\mathbf{a}_p$ of $\mathbf{a}$ won't affect the constraint in (6). Since the objective is to be minimized, we get $\mathbf{a}_p = \mathbf{0}$, so $\mathbf{a} = \mathbf{a}_d$: the optimal $\mathbf{a}$ will lie in the span of the data points $\{\mathbf{x}_i\}_{i=1}^{N_x}$ and $\{\mathbf{y}_i\}_{i=1}^{N_y}$. Notice that if $\rho_x = \rho_y = 0$, both objective and constraint of (6) will not be affected by an orthogonal component $\mathbf{a}_p$ of $\mathbf{a}$. Since $\mathbf{a}_p$ doesn't play any role at all, we can as well put $\mathbf{a}_p = 0$ and obtain the minimal norm $\mathbf{a}$. □

As a consequence, we can write $\mathbf{a}$ as a linear combination of the data points and then solve for the coefficients. By doing so, one can easily check that the optimization problem (6) can be entirely expressed in terms of inner products between data points $\{\mathbf{x}_i\}_{i=1}^{N_x}$ and $\{\mathbf{y}_i\}_{i=1}^{N_y}$ only, if the conditions of the corollary are fulfilled. This will make the kernelization of our approach possible.

5.2 Notation

Before stating the "kernelized" version of Theorem 2, we introduce the following notation. Let $\{\mathbf{x}_i\}_{i=1}^{N_x}$ and $\{\mathbf{y}_i\}_{i=1}^{N_y}$ be the training data points in the classes corresponding to $\mathbf{x}$ and $\mathbf{y}$ respectively, and let $\{\mathbf{t}_i\}_{i=1}^{N}$ denote the set of all $N = N_x + N_y$ training data points where

$$\begin{aligned} \mathbf{t}_i &= \mathbf{x}_i \qquad & i = 1, 2, \ldots, N_x \\ \mathbf{t}_i &= \mathbf{y}_{i-N_x} \qquad & i = N_x + 1, N_x + 2, \ldots, N. \end{aligned}$$

The Gram matrix $\mathbf{K}$ can now be defined as $\mathbf{K}_{ij} = \varphi(\mathbf{t}_i)^T \varphi(\mathbf{t}_j) = K(\mathbf{t}_i, \mathbf{t}_j)$ for $i, j = 1, 2, \ldots, N$ where the first N_x rows and the last N_y rows of $\mathbf{K}$ are denoted by $\mathbf{K}_\mathbf{x}$ and $\mathbf{K}_\mathbf{y}$, respectively:

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}_\mathbf{x} \\ \mathbf{K}_\mathbf{y} \end{pmatrix}.$$

The block-row-averaged Gram matrix $\mathbf{L}$ is then obtained by setting the row average of the $\mathbf{K}_\mathbf{x}$-block and the $\mathbf{K}_\mathbf{y}$-block equal to zero:

$$\mathbf{L} = \begin{pmatrix} \mathbf{K}_\mathbf{x} - \mathbf{1}_{N_x} \mathbf{l}_\mathbf{x}^T \\ \mathbf{K}_\mathbf{y} - \mathbf{1}_{N_y} \mathbf{l}_\mathbf{y}^T \end{pmatrix} = \begin{pmatrix} \sqrt{N_x}\mathbf{L}_\mathbf{x} \\ \sqrt{N_y}\mathbf{L}_\mathbf{y} \end{pmatrix},$$

where $\mathbf{1}_m$ is a column vector of ones of dimension m. The row averages $\mathbf{l}_\mathbf{x}^T$ and $\mathbf{l}_\mathbf{y}^T$ are N-dimensional row vectors given by:

$$\begin{aligned} (\mathbf{l}_\mathbf{x}^T)_i &= \frac{1}{N_x} \sum_{j=1}^{N_x} K(\mathbf{x}_j, \mathbf{t}_i), \\ (\mathbf{l}_\mathbf{y}^T)_i &= \frac{1}{N_y} \sum_{j=1}^{N_y} K(\mathbf{y}_j, \mathbf{t}_i). \end{aligned}$$

5.3 Result

Theorem 6 *If $\mathbf{l}_\mathbf{x} = \mathbf{l}_\mathbf{y}$, then the minimax probability decision problem (28) does not have a solution. Otherwise, an optimal decision boundary $\mathcal{D}(\gamma, b)$ can be determined by solving the convex optimization problem:*

$$\kappa_*^{-1} := \min_\gamma \sqrt{\gamma^T \mathbf{L}_\mathbf{x}^T \mathbf{L}_\mathbf{x} \gamma} + \sqrt{\gamma^T \mathbf{L}_\mathbf{y}^T \mathbf{L}_\mathbf{y} \gamma} \quad s.t. \quad \gamma^T(\mathbf{l}_\mathbf{x} - \mathbf{l}_\mathbf{y}) = 1, \tag{29}$$

and setting b to the value

$$b_* = \gamma_*^T \mathbf{l}_\mathbf{x} - \kappa_* \sqrt{\gamma_*^T \mathbf{L}_\mathbf{x}^T \mathbf{L}_\mathbf{x} \gamma_*},$$

where γ_* is an optimal solution of (29). The optimal worst-case misclassification probability is obtained via

$$1 - \alpha_* = \frac{1}{1 + \kappa_*^2} = \frac{\left(\sqrt{\gamma_*^T \mathbf{L}_\mathbf{x}^T \mathbf{L}_\mathbf{x} \gamma_*} + \sqrt{\gamma_*^T \mathbf{L}_\mathbf{y}^T \mathbf{L}_\mathbf{y} \gamma_*}\right)^2}{1 + \left(\sqrt{\gamma_*^T \mathbf{L}_\mathbf{x}^T \mathbf{L}_\mathbf{x} \gamma_*} + \sqrt{\gamma_*^T \mathbf{L}_\mathbf{y}^T \mathbf{L}_\mathbf{y} \gamma_*}\right)^2}.$$

Because both $\mathbf{L}_\mathbf{x}^T \mathbf{L}_\mathbf{x}$ and $\mathbf{L}_\mathbf{y}^T \mathbf{L}_\mathbf{y}$ are positive semidefinite but not positive definite, the optimal hyperplane is not unique.

Once an optimal decision boundary is found, classification of a new data point $\mathbf{z}_{new}$ is done by evaluating $\text{sign}(\mathbf{a}_*^T\varphi(\mathbf{z}_{new}) - b_*) = \text{sign}\left(\left(\sum_{i=1}^{N_x+N_y}[\gamma_*]_i K(\mathbf{t}_i, \mathbf{z}_{new})\right) - b_*\right)$ (notice that this can be evaluated only in terms of the kernel function); if this is $+1$, $\mathbf{z}_{new}$ is classified as belonging to class $\mathbf{x}$, otherwise $\mathbf{z}_{new}$ is classified as belonging to class $\mathbf{y}$.

Proof of theorem: In order to solve (28), we can apply Theorem 2. If the distributions of $\varphi(\mathbf{x})$ and $\varphi(\mathbf{y})$ have the same mean: $\overline{\varphi(\mathbf{x})} = \overline{\varphi(\mathbf{y})}$, then the minimax probability decision problem (28) does not have a solution. Otherwise, an optimal hyperplane $\mathcal{H}(\mathbf{a}, b)$ in $\mathbb{R}^f$ can be determined by solving the convex optimization problem:

$$\kappa_*^{-1} := \min_{\mathbf{a}} \sqrt{\mathbf{a}^T\Sigma_{\varphi(\mathbf{x})}\mathbf{a}} + \sqrt{\mathbf{a}^T\Sigma_{\varphi(\mathbf{y})}\mathbf{a}} \quad \text{s.t.} \quad \mathbf{a}^T(\overline{\varphi(\mathbf{x})} - \overline{\varphi(\mathbf{y})}) = 1, \tag{30}$$

and setting b to the value

$$b_* = \mathbf{a}_*^T\overline{\varphi(\mathbf{x})} - \kappa_*\sqrt{\mathbf{a}_*^T\Sigma_{\varphi(\mathbf{x})}\mathbf{a}_*}, \tag{31}$$

where $\mathbf{a}_*$ is an optimal solution of (30).

However, we do not wish to solve the convex optimization problem in this form, because we want to avoid using f or φ explicitly. We first form sample estimates of the means and covariances $\overline{\varphi(\mathbf{x})}, \overline{\varphi(\mathbf{y})}, \Sigma_{\varphi(\mathbf{x})}$ and $\Sigma_{\varphi(\mathbf{y})}$:

$$\begin{aligned}
\widehat{\varphi(\mathbf{x})} &= \frac{1}{N_x}\sum_{i=1}^{N_x}\varphi(\mathbf{x}_i), \\
\widehat{\varphi(\mathbf{y})} &= \frac{1}{N_y}\sum_{i=1}^{N_y}\varphi(\mathbf{y}_i), \\
\hat{\mathbf{\Sigma}}_{\varphi(\mathbf{x})} &= \frac{1}{N_x}\sum_{i=1}^{N_x}(\varphi(\mathbf{x}_i) - \widehat{\varphi(\mathbf{x})})(\varphi(\mathbf{x}_i) - \widehat{\varphi(\mathbf{x})})^T, \\
\hat{\mathbf{\Sigma}}_{\varphi(\mathbf{y})} &= \frac{1}{N_y}\sum_{i=1}^{N_y}(\varphi(\mathbf{y}_i) - \widehat{\varphi(\mathbf{y})})(\varphi(\mathbf{y}_i) - \widehat{\varphi(\mathbf{y})})^T,
\end{aligned}$$

and plug these into the objective function and the constraint in (30):

$$\kappa_*^{-1} := \min_{\mathbf{a}} \sqrt{\mathbf{a}^T\hat{\mathbf{\Sigma}}_{\varphi(\mathbf{x})}\mathbf{a}} + \sqrt{\mathbf{a}^T\hat{\mathbf{\Sigma}}_{\varphi(\mathbf{y})}\mathbf{a}} \quad \text{s.t.} \quad \mathbf{a}^T(\widehat{\varphi(\mathbf{x})} - \widehat{\varphi(\mathbf{y})}) = 1. \tag{32}$$

Those estimates satisfy the conditions of Corollary 5, now applied in $\mathbb{R}^f$. So, without loss of generality, we can write $\mathbf{a}$ as

$$\mathbf{a} = \sum_{i=1}^{N_x}\alpha_i\varphi(\mathbf{x}_i) + \sum_{j=1}^{N_y}\beta_j\varphi(\mathbf{y}_j). \tag{33}$$

This last equation could be derived in a more formal way by using the representer theorem (Schölkopf and Smola, 2002).

As explained above, a consequence of Corollary 5 is that objective and constraints can be entirely expressed in terms of inner products of $\varphi(\mathbf{z})$ only. When substituting expression (33) for

$\mathbf{a}$ in (32), we indeed see that both the objective and the constraints can be written in terms of the kernel function $K(\mathbf{z}_1, \mathbf{z}_2) = \varphi(\mathbf{z}_1)^T \varphi(\mathbf{z}_2)$. We obtain:

$$\kappa_*^{-1} := \min_\gamma \sqrt{\gamma^T \mathbf{L}_\mathbf{x}^T \mathbf{L}_\mathbf{x} \gamma} + \sqrt{\gamma^T \mathbf{L}_\mathbf{y}^T \mathbf{L}_\mathbf{y} \gamma} \quad \text{s.t.} \quad \gamma^T (\mathbf{l}_\mathbf{x} - \mathbf{l}_\mathbf{y}) = 1,$$

where $\gamma = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_{N_x} \ \beta_1 \ \beta_2 \ \cdots \ \beta_{N_y}]^T$. We can also write this as

$$\kappa_*^{-1} := \min_\gamma \|\mathbf{L}_\mathbf{x} \gamma\|_2 + \|\mathbf{L}_\mathbf{y} \gamma\|_2 \quad \text{s.t.} \quad \gamma^T (\mathbf{l}_\mathbf{x} - \mathbf{l}_\mathbf{y}) = 1,$$

which is a second order cone program (SOCP) (Boyd and Vandenberghe, 2001) that has the same form as the SOCP in (12) and can thus be solved in a similar way. Notice that, in this case, the optimizing variable is $\gamma \in \mathbb{R}^{N_x + N_y}$ instead of $\mathbf{a} \in \mathbb{R}^n$. Thus the dimension of the optimization problem increases, but the solution is more powerful because the kernelization corresponds to a more complex decision boundary in $\mathbb{R}^n$.

Similarly, the optimal value b_* of b in (31) will then become

$$b_* = \gamma_*^T \mathbf{l}_\mathbf{x} - \kappa_* \sqrt{\gamma_*^T \mathbf{L}_\mathbf{x}^T \mathbf{L}_\mathbf{x} \gamma_*} = \gamma_*^T \mathbf{l}_\mathbf{y} + \kappa_* \sqrt{\gamma_*^T \mathbf{L}_\mathbf{y}^T \mathbf{L}_\mathbf{y} \gamma_*},$$

where γ_* is an optimal value of γ. □

Remark. Notice that this kernelization will in general be possible when expressing $\overline{\varphi(\mathbf{x})}, \overline{\varphi(\mathbf{y})}, \mathbf{\Sigma}_{\varphi(\mathbf{x})}$, and $\mathbf{\Sigma}_{\varphi(\mathbf{y})}$ as

$$\begin{aligned}
\overline{\varphi(\mathbf{x})} &= \sum_{i=1}^{N_x} \lambda_i \varphi(\mathbf{x}_i), \\
\overline{\varphi(\mathbf{y})} &= \sum_{i=1}^{N_y} \omega_i \varphi(\mathbf{y}_i), \\
\mathbf{\Sigma}_{\varphi(\mathbf{x})} &= \rho_x I_f + \sum_{i=1}^{N_x} \Lambda_i (\varphi(\mathbf{x}_i) - \overline{\varphi(\mathbf{x})})(\varphi(\mathbf{x}_i) - \overline{\varphi(\mathbf{x})})^T, \\
\mathbf{\Sigma}_{\varphi(\mathbf{x})} &= \rho_y I_f + \sum_{i=1}^{N_y} \Omega_i (\varphi(\mathbf{y}_i) - \overline{\varphi(\mathbf{y})})(\varphi(\mathbf{y}_i) - \overline{\varphi(\mathbf{y})})^T,
\end{aligned}$$

where I_f is the identity matrix of dimension f. Indeed, from Corollary 5, we know that the optimal $\mathbf{a}$ will lie in the span of the mapped data $\{\varphi(\mathbf{t}_i)\}_{i=1}^N$ and as a consequence, the optimization problem (28) can be entirely expressed in terms of inner products of the mapped data only. Even if we take uncertainty in the means and covariance matrices into account (that is $\nu \neq 0, \rho_x = \rho_y \neq 0$, see Section 3), we can still kernelize the MPM and obtain the convex optimization problem

$$\kappa_*^{-1} := \min_\gamma \sqrt{\gamma^T (\mathbf{L}_\mathbf{x}^T \mathbf{L}_\mathbf{x} + \rho_x \mathbf{K}) \gamma} + \sqrt{\gamma^T (\mathbf{L}_\mathbf{y}^T \mathbf{L}_\mathbf{y} + \rho_y \mathbf{K}) \gamma} \quad \text{s.t.} \quad \gamma^T (\mathbf{l}_\mathbf{x} - \mathbf{l}_\mathbf{y}) = 1,$$

and optimal value for b

$$b_* = \gamma_*^T \mathbf{l}_\mathbf{x} - \kappa_* \sqrt{\gamma_*^T (\mathbf{L}_\mathbf{x}^T \mathbf{L}_\mathbf{x} + \rho_x \mathbf{K}) \gamma_*}.$$

The optimal worst-case misclassification probability is obtained via

$$1 - \alpha_*^{\text{rob}} = \frac{1}{1 + \max(0, (\kappa_* - \nu))^2}.$$

This approach is able to deal with uncertainty in the means and covariance matrices of the mapped data. However, it will not necessarily regularize the kernel MPM, since the Gram matrix $\mathbf{K}$ may not be positive definite. For regularization purposes, it is thus advantageous to add a term ρI_N to $\mathbf{K}$, where $\rho > 0$ is small.

5.4 Single Class Case

In a similar way, a nonlinear region $\mathcal{Q}$ for the single class case can be obtained in $\mathbb{R}^n$ by mapping the data $\mathbf{x} \mapsto \varphi(\mathbf{x}) \sim (\overline{\varphi(\mathbf{x})}, \mathbf{\Sigma}_{\varphi(\mathbf{x})})$ and solving (25) in the higher-dimensional feature space $\mathbb{R}^f$:

$$\max_{\mathbf{a}\neq 0, b} \frac{b}{\sqrt{\mathbf{a}^T \mathbf{\Sigma}_{\varphi(\mathbf{x})} \mathbf{a}}} \quad \text{s.t.} \quad \inf_{\varphi(\mathbf{x}) \sim (\overline{\varphi(\mathbf{x})}, \mathbf{\Sigma}_{\varphi(\mathbf{x})})} \mathbf{Pr}\{\mathbf{a}^T \varphi(\mathbf{x}) \geq b\} \geq \alpha. \tag{34}$$

Again, this optimization problem can be reformulated in terms of a given kernel function $K(\mathbf{z}_1, \mathbf{z}_2) = \varphi(\mathbf{z}_1)^T \varphi(\mathbf{z}_2)$ satisfying Mercer's condition.

Theorem 7 *Let $\alpha \in (0, 1]$. If this choice of α is feasible, that is*

$$\exists\, \gamma \ : \ \gamma^T \mathbf{l}_\mathbf{x} - 1 \geq \kappa(\alpha) \sqrt{\gamma^T \mathbf{L}_\mathbf{x}^T \mathbf{L}_\mathbf{x} \gamma},$$

then an optimal region $\mathcal{Q}(\gamma, b)$ can be determined by solving the (convex) second order cone programming problem:

$$\min_{\gamma} \ \gamma^T \mathbf{L}_\mathbf{x}^T \mathbf{L}_\mathbf{x} \gamma \quad s.t. \quad \gamma^T \mathbf{l}_\mathbf{x} - 1 \geq \kappa(\alpha) \sqrt{\gamma^T \mathbf{L}_\mathbf{x}^T \mathbf{L}_\mathbf{x} \gamma},$$

where $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$ and $b = 1$. The worst-case probability of occurrence outside region $\mathcal{Q}$ is given by $1 - \alpha$. If $\mathbf{L}_\mathbf{x}^T \mathbf{L}_\mathbf{x}$ is positive definite, the optimal half-space is unique and determined by

$$\gamma_* = \frac{(\mathbf{L}_\mathbf{x}^T \mathbf{L}_\mathbf{x})^{-1} \mathbf{l}_\mathbf{x}}{\zeta^2 - \kappa(\alpha)\zeta} \qquad \text{with} \quad \zeta = \sqrt{\mathbf{l}_\mathbf{x}^T (\mathbf{L}_\mathbf{x}^T \mathbf{L}_\mathbf{x})^{-1} \mathbf{l}_\mathbf{x}},$$

if the choice of α is such that $\kappa(\alpha) \leq \zeta$ or $\alpha \leq \frac{\zeta^2}{1+\zeta^2}$.

If the choice of α is not feasible or if $\mathbf{l}_\mathbf{x} = \mathbf{0}$ (in this case, no $\alpha \in (0, 1]$ will be feasible), problem (34) does not have a solution.

Once an optimal decision region is found, future points $\mathbf{z}$ for which $\mathbf{a}_*^T \varphi(\mathbf{z}) = \sum_{i=1}^{N_x} [\gamma_*]_i K(\mathbf{x}_i, \mathbf{z}) \leq b_*$ (notice that this can be evaluated only in terms of the kernel function), can then be considered as outliers with respect to the region $\mathcal{Q}$, with the worst-case probability of occurrence outside $\mathcal{Q}$ given by $1 - \alpha$.

6 Experiments

In this section we report the results of experiments that we carried out to test our algorithmic approach. The validity of $1-\alpha$ as the worst-case probability of misclassification of future data is checked, and we also assess the usefulness of the kernel trick in this setting. We compare linear kernels and Gaussian kernels.

Experimental results on standard benchmark problems are summarized in Table 2. The Wisconsin breast cancer dataset contains 16 missing examples which were not used. The breast cancer, pima diabetes, ionosphere and sonar data were obtained from the UCI repository. The pima diabetes data were normalized. Data for the twonorm problem were generated as specified by Breiman (1998). Each dataset was randomly partitioned into 90% training and 10% test sets. The kernel parameter (σ) for the Gaussian kernel ($e^{-\|x-y\|^2/\sigma}$) was tuned using cross-validation over 50 random partitions of the training set. The reported results are the averages over 50 random partitions for both the linear kernel and the Gaussian kernel with σ chosen as above.

Dataset	*Linear kernel*		*Gaussian kernel*		*BPB*	*SVML*	*SVMG*
	α	TSA	α	TSA			
Twonorm	80.4±0.1 %	95.8±0.4 %	91.3±0.1 %	95.7±0.5 %	96.3 %	95.1±0.6 %	96.1±0.4 %
Breast cancer	84.4±0.1 %	97.0±0.4 %	89.1 ± 0.1 %	96.9±0.3 %	96.8 %	96.4±0.4 %	96.5 ± 0.3 %
Ionosphere	65.5±0.3 %	83.4±0.9 %	89.3±0.2 %	91.5±0.7 %	93.7 %	87.1±0.9 %	94.1±0.7 %
Pima diabetes	32.2±0.2 %	76.3±0.6 %	32.5±0.2 %	76.2±0.6 %	76.1 %	77.9±0.7 %	77.9±0.7 %
Sonar	67.0±0.4 %	74.9±1.4 %	99.9 ± 0.1 %	87.5±0.9 %	-	76.1±1.5 %	86.6±1.0 %

Table 2: Lower bound α and test-set accuracy (TSA) compared to BPB (best performance in the article of Breiman (1998)) and to the performance of an SVM with linear kernel (SVML) and an SVM with Gaussian kernel (SVMG).

The results are comparable with those in the existing literature (Breiman, 1998) and with those obtained with support vector machines. Also, we notice that α is smaller than the test-set accuracy in all cases, except for the sonar data, with Gaussian kernel. Furthermore, α is smaller for a linear decision boundary then for the nonlinear decision boundary obtained via the Gaussian kernel. This clearly shows that kernelizing the method leads to more powerful decision boundaries.

We notice that, for the above binary classification problems with benchmark data, using plug-in estimates of mean and covariance matrix without robustness considerations is successful—in some sense the bias incurred by the use of plug-in estimates in the two classes appears to "cancel" and have diminished overall impact on the discriminant boundary. To show that having poor plug-in estimates for one class and better plug-in estimates for the other can bias the solution and the lower bound α, and that a robust approach can improve this result, consider the following toy example (Matlab code for this example is available on *http://robotics.eecs.berkeley.edu/˜gert/*).

Suppose data points of class $\mathbf{x}$ are generated by a 2-dimensional Gaussian distribution with $\bar{\mathbf{x}} = [0\ 0]^T$ and $\mathbf{\Sigma_x} = I$ and data points of class $\mathbf{y}$ by another 2-dimensional Gaussian distribution with $\bar{\mathbf{y}} = [0\ 3]^T$ and $\mathbf{\Sigma_y} = I$. A small training data sample consisting of 10 points of each class is depicted in Figure 3. As one can see, the few training samples from class $\mathbf{x}$ are concentrated around the mean, while the few training samples from class $\mathbf{y}$ are somewhat more spread out. The empirical estimates $\hat{\mathbf{\Sigma}}_\mathbf{x}$ and $\hat{\mathbf{\Sigma}}_\mathbf{y}$ of $\mathbf{\Sigma_x}$ and $\mathbf{\Sigma_y}$ will reflect this imbalance, leading to the false impression of a more concentrated distribution for class $\mathbf{x}$ than for class $\mathbf{y}$. As expected, training a standard linear MPM will lead to a decision line clearly biased towards the origin, as shown by the dotted red line in Figure 3. The lower bound on the correct classification of future data points (assuming Gaussian distributions) turns out to be 95.6%. However, for a test set consisting of 50 points from each class (see Figure 3), the obtained accuracy is only 90%, which obviously violates the lower bound.

To see how a robust linear approach improves this result, we train a robust linear MPM specifying the uncertainty parameters as $\rho_x = \|\mathbf{\Sigma_x} - \hat{\mathbf{\Sigma}}_\mathbf{x}\|_F = 0.87$, $\rho_y = \|\mathbf{\Sigma_y} - \hat{\mathbf{\Sigma}}_\mathbf{y}\|_F = 0.19$ and $\nu^2 = \max\{(\bar{\mathbf{x}} - \hat{\mathbf{x}})^T \mathbf{\Sigma_x}^{-1}(\bar{\mathbf{x}} - \hat{\mathbf{x}}), (\bar{\mathbf{y}} - \hat{\mathbf{y}})^T \mathbf{\Sigma_y}^{-1}(\bar{\mathbf{y}} - \hat{\mathbf{y}})\} = 0.59$, using our knowledge of the real means and covariance matrices in this case. One can notice that indeed $\rho_x > \rho_y$, corresponding to the clear difference in quality of the empirical moment estimates for both classes, as noticed earlier. Moreover, the robust approach clearly reduces the bias on the decision line as can be seen in Figure 3 (solid red line). The lower bound on the correct classification of future data points decreases to 76.8%, while the performance on the same test data increases to 94%. So, the lower bound is respected in the robust case and the performance improves. For comparison, the dashed black line in Figure 3 corresponds to a linear 1-norm soft margin SVM with regularization parameter $C = 1$, with a test set performance of 92%.

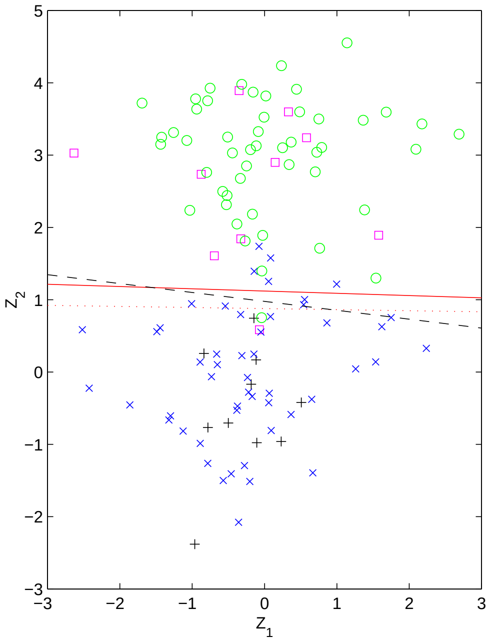


Figure 3: Minimax probabilistic decision line (dotted red line), robust minimax probabilistic decision line (solid red line) and the decision line for a linear 1-norm soft margin SVM with $C = 1$ (dashed black line) in $\mathbb{R}^2$. Training points are indicated with black $+$'s for class $\mathbf{x}$ and magenta $\square$'s for class $\mathbf{y}$. Test points are indicated with blue $\times$'s for class $\mathbf{x}$ and green o's for class $\mathbf{y}$. Notice how the robust approach clearly reduces the bias on the decision line and improves the test set performance compared to the standard linear MPM.

This shows how the robust MPM clearly improves on the standard MPM when a bias is incurred by the use of plug-in estimates of unbalanced quality in the two classes. For most of the benchmark data sets used for Table 2, this imbalance doesn't seem to be significant. However, in the single-class setting, the uncertainty due to estimation of $\bar{\mathbf{x}}$ and $\mathbf{\Sigma}_{\mathbf{x}}$ translates directly into movement of the discriminant boundary and cannot be neglected. For the robust non-linear (kernelized) version of the single-class MPM, as well as experimental results in robust novelty detection, we refer to Lanckriet et al. (2002b).

7 Conclusions

The problem of linear discrimination has a long and distinguished history. Many results on misclassification rates have been obtained by making distributional assumptions (e.g., Anderson and Bahadur, 1962). Our results, on the other hand, make use of the moment-based inequalities of Marshall and Olkin (1960) to obtain distribution-free results for linear discriminants. We considered the case of binary classification, where only the mean and covariance matrix of the classes are assumed to be known. The minimax probabilistic decision hyperplane is then determined by optimizing the worst-case probabilities over all possible class-conditional distributions.

We have addressed issues of robustness when plug-in estimates are used for means and covariance matrices instead of their real, but unknown values. We have also shown how to exploit Mercer kernels to generalize our algorithm to nonlinear classification. Experimental results show that our method is competitive with the existing literature and furthermore, our method yields an explicit upper bound on the probability of future misclassification.

The computational complexity of our method is comparable to the quadratic program that one has to solve for the support vector machine (SVM). While we have used a simple iterative least-squares approach, we believe that there is much to gain from exploiting analogies to the SVM and developing specialized optimization procedures for our algorithm, in particular procedures that break the data into subsets. Another direction that we are currently investigating is the extension of our approach to multiway classification.

Acknowledgments

We would like to acknowledge support from ONR MURI N00014-00-1-0637, from NSF grants IIS-9988642 and ECS-9983874, from the Belgian American Educational Foundation and from DARPA grant F33615-01-C-3150. We also thank Alex Smola for constructive conversations and suggestions.

A Proof of Expression (5)

We want to find a closed-form expression for d^2:

$$d^2 = \inf_{\mathbf{a}^T\mathbf{y} \geq b} (\mathbf{y} - \bar{\mathbf{y}})^T \mathbf{\Sigma}_{\mathbf{y}}{}^{-1} (\mathbf{y} - \bar{\mathbf{y}}).$$

First notice that if $\mathbf{a}^T\bar{\mathbf{y}} \geq b$, then we can just take $\mathbf{y} = \bar{\mathbf{y}}$ and obtain $d^2 = 0$, which is certainly the optimum because $d^2 \geq 0$ and we take the infimum.

Let's now assume $\mathbf{a}^T\bar{\mathbf{y}} \leq b$. We can write this as $d^2 = \inf_{\mathbf{c}^T\mathbf{w}\geq f}\ \mathbf{w}^T\mathbf{w}$, where $\mathbf{w} = {\boldsymbol{\Sigma}_\mathbf{y}}^{-1/2}(\mathbf{y}-\bar{\mathbf{y}})$, $\mathbf{c}^T = \mathbf{a}^T{\boldsymbol{\Sigma}_\mathbf{y}}^{1/2}$ and $f = b - \mathbf{a}^T\bar{\mathbf{y}} \geq 0$. We then form the Lagrangian:

$$\mathcal{L}(\mathbf{w}, \lambda) = \mathbf{w}^T\mathbf{w} + \lambda(f - \mathbf{c}^T\mathbf{w}),$$

which is to be maximized with respect to $\lambda \geq 0$ and minimized with respect to $\mathbf{w}$. At the optimum, $2\mathbf{w} = \lambda\mathbf{c}$ and $f = \mathbf{c}^T\mathbf{w}$. So, $\lambda = \frac{2f}{\mathbf{c}^T\mathbf{c}}$ such that indeed $\lambda \geq 0$ because $f > 0$. Also, $\mathbf{w} = \frac{d\mathbf{c}}{\mathbf{c}^T\mathbf{c}}$. This yields:

$$d^2 = \inf_{\mathbf{a}^T\mathbf{y}\geq b} (\mathbf{y}-\bar{\mathbf{y}})^T{\boldsymbol{\Sigma}_\mathbf{y}}^{-1}(\mathbf{y}-\bar{\mathbf{y}}) = \frac{(b - \mathbf{a}^T\bar{\mathbf{y}})^2}{\mathbf{a}^T\boldsymbol{\Sigma}_\mathbf{y}\mathbf{a}}.$$

Combining both cases $\mathbf{a}^T\bar{\mathbf{y}} \geq b$ and $\mathbf{a}^T\bar{\mathbf{y}} \leq b$ in one expression:

$$d^2 = \inf_{\mathbf{a}^T\mathbf{y}\geq b} (\mathbf{y}-\bar{\mathbf{y}})^T{\boldsymbol{\Sigma}_\mathbf{y}}^{-1}(\mathbf{y}-\bar{\mathbf{y}}) = \frac{\max((b - \mathbf{a}^T\bar{\mathbf{y}}), 0)^2}{\mathbf{a}^T\boldsymbol{\Sigma}_\mathbf{y}\mathbf{a}},$$

which proves expression (5).

B Proof of Expression (21)

To solve the constrained optimization problem

$$\min_{\bar{\mathbf{x}}\,:\,(\bar{\mathbf{x}}-\bar{\mathbf{x}}^0)^T{\boldsymbol{\Sigma}_\mathbf{x}}^{-1}(\bar{\mathbf{x}}-\bar{\mathbf{x}}^0)\leq\nu^2} \mathbf{a}^T\bar{\mathbf{x}},$$

we form the Lagrangian

$$\mathcal{L}(\bar{\mathbf{x}}, \lambda) = \mathbf{a}^T\bar{\mathbf{x}} + \lambda((\bar{\mathbf{x}}-\bar{\mathbf{x}}^0)^T{\boldsymbol{\Sigma}_\mathbf{x}}^{-1}(\bar{\mathbf{x}}-\bar{\mathbf{x}}^0) - \nu^2),$$

which is to be maximized with respect to $\lambda \geq 0$ and minimized with respect to $\bar{\mathbf{x}}$. At the optimum:

$$\begin{aligned}
\frac{\partial}{\partial\bar{\mathbf{x}}}\mathcal{L}(\bar{\mathbf{x}}, \lambda) &= \mathbf{a} + 2\lambda{\boldsymbol{\Sigma}_\mathbf{x}}^{-1}\bar{\mathbf{x}} - 2\lambda{\boldsymbol{\Sigma}_\mathbf{x}}^{-1}\bar{\mathbf{x}}^0 = 0 \\
\Rightarrow \bar{\mathbf{x}} &= \bar{\mathbf{x}}^0 - \frac{1}{2\lambda}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{a}, &(35)\\
\frac{\partial}{\partial\lambda}\mathcal{L}(\bar{\mathbf{x}}, \lambda) &= (\bar{\mathbf{x}}-\bar{\mathbf{x}}^0)^T{\boldsymbol{\Sigma}_\mathbf{x}}^{-1}(\bar{\mathbf{x}}-\bar{\mathbf{x}}^0) - \nu^2 \\
&= \frac{1}{4\lambda^2}\mathbf{a}^T\boldsymbol{\Sigma}_\mathbf{x}\mathbf{a} - \nu^2 = 0 &(36)\\
\Rightarrow \lambda &= \sqrt{\frac{\mathbf{a}^T\boldsymbol{\Sigma}_\mathbf{x}\mathbf{a}}{4\nu}}, &(37)
\end{aligned}$$

where (35) is substituted to obtain (36). Substitution of (37) in (35) gives the optimal value for $\bar{\mathbf{x}}$ and leads to:

$$\min_{\bar{\mathbf{x}}\,:\,(\bar{\mathbf{x}}-\bar{\mathbf{x}}^0)^T{\boldsymbol{\Sigma}_\mathbf{x}}^{-1}(\bar{\mathbf{x}}-\bar{\mathbf{x}}^0)\leq\nu^2} \mathbf{a}^T\bar{\mathbf{x}} = \mathbf{a}^T\bar{\mathbf{x}}^0 - \nu\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_\mathbf{x}\mathbf{a}},$$

which proves expression (21).

C Proof of Expression (23)

To solve the constrained optimization problem

$$\max_{\mathbf{\Sigma}} \mathbf{a}^T \mathbf{\Sigma} \mathbf{a} \; : \; \|\mathbf{\Sigma} - \mathbf{\Sigma}^0\|_F \leq \rho,$$

we can let $\mathbf{\Sigma}$ be of the form $\mathbf{\Sigma} = \mathbf{\Sigma}^0 + \rho \Delta \mathbf{\Sigma}$, without loss of generality. Recall that the norm used here is the Frobenius norm: $\|A\|_F^2 = \mathbf{Tr}(A^T A)$. The optimization problem then becomes:

$$\max_{\Delta \mathbf{\Sigma} \; : \; \|\Delta \mathbf{\Sigma}\|_F \leq 1} \mathbf{a}^T \mathbf{\Sigma}^0 \mathbf{a} + \rho \mathbf{a}^T \Delta \mathbf{\Sigma} \mathbf{a}.$$

To find the value of $\Delta \mathbf{\Sigma}$ that maximizes this, we first solve

$$\max_{\Delta \mathbf{\Sigma} \; : \; \|\Delta \mathbf{\Sigma}\|_F \leq 1} \mathbf{a}^T \Delta \mathbf{\Sigma} \mathbf{a}. \tag{38}$$

Notice that, using the Cauchy-Schwarz inequality, we have

$$\mathbf{a}^T \Delta \mathbf{\Sigma} \mathbf{a} \leq \|\mathbf{a}\|_2 \|\Delta \mathbf{\Sigma} \mathbf{a}\|_2 \leq \|\mathbf{a}\|_2 \|\Delta \mathbf{\Sigma}\|_F \|\mathbf{a}\|_2 \leq \|\mathbf{a}\|_2^2,$$

because of compatibility of the Frobenius matrix norm and the euclidean vector norm and because $\|\Delta \mathbf{\Sigma}\|_F \leq 1$.

So, the objective of (38) is upper bounded by $\|\mathbf{a}\|_2^2$. For $\Delta \mathbf{\Sigma}$ the unity matrix, this bound is attained. We obtain

$$\max_{\mathbf{\Sigma} \; : \; \|\mathbf{\Sigma} - \mathbf{\Sigma}^0\|_F \leq \rho} \mathbf{a}^T \mathbf{\Sigma} \mathbf{a} = \mathbf{a}^T \left(\mathbf{\Sigma}^0 + \rho I_n \right) \mathbf{a} = \mathbf{a}^T \mathbf{\Sigma}^0 \mathbf{a} + \rho \mathbf{a}^T \mathbf{a},$$

which proves expression (23).

References

Andersen, E. D. and Andersen, A. D. (2000). The MOSEK interior point optimizer for linear programming: An implementation of the homogeneous algorithm. In Frenk, H., Roos, C., Terlaky, T., and Zhang, S., editors, *High Performance Optimization*, pages 197–232. Kluwer Academic Publishers.

Anderson, T. W. and Bahadur, R. R. (1962). Classification into two multivariate Normal distributions with different covariance matrices. *Annals of Mathematical Statistics*, 33(2):420–431.

Bennett, K. P. and Bredensteiner, E. J. (2000). Duality and geometry in SVM classifiers. In *Proc. 17th International Conf. on Machine Learning*, pages 57–64. Morgan Kaufmann, San Francisco, CA.

Bertsekas, D. P. (1999). *Nonlinear Programming.* Athena Scientific, Belmont, MA.

Boyd, S. and Vandenberghe, L. (2001). Convex optimization. Course notes for EE364, Stanford University. Available at `http://www.stanford.edu/class/ee364`.

Breiman, L. (1998). Arcing classifiers. *Annals of Statistics*, 26(3):801–849.

Chapelle, O., Weston, J., Bottou, L., and Vapnik, V. (2001). Vicinal risk minimization. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 416–422. MIT Press.

Crisp, D. and Burges, C. (1999). A geometric interpretation of ν-SVM classifiers. In Solla, S., Leen, T., and Muller, K., editors, *Advances in Neural Information Processing Systems 12*, pages 244–251. MIT Press, Cambridge, MA.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition.* Academic Press, San Diego, CA, 2nd edition.

Kass, R., Tierney, L., and Kadane, J. (1988). Asymptotics in Bayesian computation. In Bernardo, J., Groot, M. D., Lindley, D., and Smith, A., editors, *Bayesian Statistics 3*, pages 261–278. Oxford University Press, Cambridge, MA.

Lanckriet, G. R. G., El Ghaoui, L., Bhattacharyya, C., and Jordan, M. I. (2002a). Minimax probability machine. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA. MIT Press.

Lanckriet, G. R. G., El Ghaoui, L., and Jordan, M. I. (2002b). Robust novelty detection with single-class MPM. In *Advances in Neural Information Processing Systems 15*, Cambridge, MA. MIT Press.

Lobo, M., Vandenberghe, L., Boyd, S., and Lebret, H. (1998). Applications of second order cone programming. *Linear Algebra and its Applications*, 284:193–228.

Marshall, A. W. and Olkin, I. (1960). Multivariate Chebyshev inequalities. *Annals of Mathematical Statistics*, 31(4):1001–1014.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., and Müller, K.-R. (1999). Fisher discriminant analysis with kernels. In Hu, Y.-H., Larsen, J., Wilson, E., and Douglas, S., editors, *Neural Networks for Signal Processing IX*, pages 41–48. IEEE.

Nesterov, Y. and Nemirovsky, A. (1994). *Interior point polynomial methods in convex programming: Theory and applications.* SIAM, Philadelphia, PA.

Pérez-Cruz, F., Alarcón-Diana, P. L., Navia-Vázquez, A., and Artés-Rodríguez, A. (2001). Fast training of support vector classifiers. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 734–740. MIT Press.

Popescu, I. and Bertsimas, D. (2001). Optimal inequalities in probability theory: A convex optimization approach. Technical Report TM62, INSEAD.

Schölkopf, B. and Smola, A. (2002). *Learning with Kernels.* MIT Press, Cambridge, MA.

Sturm, J. (1999). Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11–12:625–653. Special issue on Interior Point Methods (CD supplement with software).

Suykens, J. A. K. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300.

Tipping, M. E. (2000). The relevance vector machine. In Solla, S., Leen, T. K., and Muller, K.-R., editors, *Advances in Neural Information Processing Systems 12*. Morgan Kaufmann.

Vapnik, V. N. (1999). *The Nature of Statistical Learning Theory*. Springer, 2nd edition.