Second Order Cone Programming formulations for feature selection

Chiranjib Bhattacharyya CHIRU@CSA.IISC.ERNET.IN
Department of Computer Science and Automation
Indian Institute of Science
Bangalore, 560 012, India

Editor: John Shawe-Taylor

Abstract

This paper addresses the issue of feature selection for linear classifiers given the moments of the class conditional densities. The problem is posed as finding a minimal set of features such that the resulting classifier has a low misclassification error. Using a bound on the misclassification error, involving the mean and covariance of class conditional densities, and minimizing a L_1 norm as an approximate criterion for feature selection a second order programming formulation is derived. To handle errors in estimation of mean and covariances a tractable robust formulation is also discussed. In a slightly different setting the Fisher discriminant is derived. Feature selection for Fisher discriminant is also discussed. Experimental results on synthetic datasets and on real life microarray data show that the proposed formulations are competitive with the state of the art linear programming formulation.

1. Introduction

The choice of useful features for discriminating between two classes is an important problem and has many applications. This paper addresses the issue of constructing linear classifiers using a small number of features when data is summarized by its moments

A linear two-class classifier is a function defined as follows

$$f(\mathbf{x}) = \mathrm{sgn}(\mathbf{w}^\top \mathbf{x} - b). \tag{1}$$

The classifier outputs 1 if the observation $\mathbf{x} \in \mathbb{R}^n$ falls in the halfspace $\{\mathbf{x}|\mathbf{w}^\top\mathbf{x} > b\}$, otherwise it outputs -1. During training the parameters, $\{\mathbf{w}, b\}$, of the discriminating hyperplane $H = \{\mathbf{x}|\mathbf{w}^\top\mathbf{x} = b\}$ is computed from a specified dataset $D = \{(\mathbf{x}_i, y_i)|\mathbf{x}_i \in \mathbb{R}^n, y_i = \{1, -1\}, i = 1, \ldots, m\}$.

Finding useful features for linear classifiers is equivalent to searching for a $\mathbf{w}$, such that most elements of $\mathbf{w}$ are zero. This can be understood as follows that if the ith component of $\mathbf{w}$ is zero, then by (1) the ith component of the observation vector $\mathbf{x}$ is irrelevant in deciding the class of $\mathbf{x}$. Using the L_0 norm of $\mathbf{w}$, defined as follows,

$$\|\mathbf{w}\|_0 = |S| \quad S = \{i|\mathbf{w}_i \neq 0\},$$

the problem of feature selection can be posed as a combinatorial optimization problem

$$\begin{array}{ll} \underset{\mathbf{w},b}{\text{minimize}} & \|\mathbf{w}\|_0 \\ \text{subject to} & y_i\left(\mathbf{w}^\top\mathbf{x}_i - b\right) \geq 1 \quad \forall 1 \leq i \leq m. \end{array} \tag{2}$$

The constraints ensure that the classifier correctly assigns labels to all training datapoints. Due to the unwieldy objective the formulation is intractable for large n (Amaldi and Kann, 1998). A heuristic tractable approximation to the proposed objective is to minimize the L_1 norm of $\mathbf{w}$. For a discussion of this issue see Chen et al. (1999) also see Weston et al. (2003) for other approximations to L_0 norm. In the sequel we will enforce the feature selection criterion by minimizing the L_1 norm.

Let $\mathbf{X}_1$ and $\mathbf{X}_2$ denote n dimensional random vectors belonging to class 1 and class 2 respectively. Without loss of generality assume that class 1 is identified with the label $y = 1$, while class 2 is identified with label $y = -1$. Let the mean and covariance of $\mathbf{X}_1$ be $\mu_1 \in \mathbb{R}^n$ and $\Sigma_1 \in \mathbb{R}^{n \times n}$ respectively. Similarly for $\mathbf{X}_2$ the mean and covariance be $\mu_2 \in \mathbb{R}^n$, and $\Sigma_2 \in \mathbb{R}^{n \times n}$ respectively. Note that Σ_1, Σ_2 are positive semidefinite symmetric matrices. In this paper we wish to address the problem of feature selection for linear classifiers given $\mu_1, \mu_2, \Sigma_1, \Sigma_2$.

Previously Lanckriet et al. (2002a,b) addressed the problem of classification given μ_1, μ_2, Σ_1, Σ_2 in a minimax setting. In their approach a chebychev inequality is used to bound the error of misclassification. We wish to use the same inequality along with the L_1 norm minimization criterion for feature selection. This leads to a Second Order Cone Programming problem (SOCP). SOCPs are a special class of nonlinear convex optimization problems, which can be efficiently solved by interior point codes (Lobo et al., 1998). We also investigate a tractable robust formulation which takes into account errors in estimating the moments.

The paper is organized as follows. In Section 2 the linear programming approach is discussed. The main contributions are in Section 3 and Section 4. Using the chebychev bound and the feature selection criterion leads to a SOCP. The Fisher discriminant is also rederived using the chebychev bound. We also discuss feature selection for the fisher discriminant. A robust formulation is discussed in section 4. Experimental results for these formulations are shown in Section 5. The concluding section summarizes the main contributions and future directions.

2. Linear Programming formulation for feature selection

The problem of finding a $\{\mathbf{w}, b\}$, such that the hyperplane $\mathbf{w}^\top \mathbf{x} = b$ discriminates well between two classes and also selects a small number of features, can be posed by the following optimization problem.

$$\begin{array}{ll} \underset{\mathbf{w},b}{\text{minimize}} & \|\mathbf{w}\|_1 \\ \text{subject to} & y_i \left(\mathbf{w}^\top \mathbf{x}_i - b\right) \geq 1 \quad \forall 1 \leq i \leq m \end{array} \tag{3}$$

At optimality it is hoped that most of the elements of the weight vector $\mathbf{w}$ are zero. The nonlinear objective is made linear by introducing two vectors such that (see Fletcher, 1989)

$$\mathbf{w} = \mathbf{u} - \mathbf{v} \quad \|\mathbf{w}\|_1 = (\mathbf{u} + \mathbf{v})^\top \mathbf{e} \quad \mathbf{u} \geq 0 \quad \mathbf{v} \geq 0 \tag{4}$$

This leads to the following Linear Programming(LP) formulation.

$$\begin{array}{ll} \underset{\mathbf{u},\mathbf{v},b}{\text{minimize}} & (\mathbf{u}+\mathbf{v})^\top \mathbf{e}, \\ \text{subject to} & y_i \left((\mathbf{u}-\mathbf{v})^\top \mathbf{x}_i - b\right) \geq 1 \quad \forall 1 \leq i \leq m \\ & \mathbf{u} \geq 0 \; \mathbf{v} \geq 0 \end{array} \tag{5}$$

The computational advantages of solving LPs make the above formulation extremely attractive. In the next section we discuss the problem of feature selection when data is summarized by the moments.

3. Feature selection using moments

Let the data for each class be specified by the first two moments, the mean and covariance. We wish to address the issue of feature selection in such a scenario. The problem is approached in a worst case setting by using a multivariate generalization of Chebychev-Cantelli inequality. The inequality is used to derive a SOCP which yields a classifier using a very small number of features.

The following multivariate generalization of Chebychev-Cantelli inequality will be used in the sequel to derive a lower bound on the probability of a random vector taking values in a given half space.

Theorem 1 *Let* $\mathbf{X}$ *be a* n *dimensional random vector. The mean and covariance of* $\mathbf{X}$ *be* $\mu \in \mathbb{R}^n$ *and* $\Sigma \in \mathbb{R}^{n \times n}$. *Let* $\mathcal{H}(\mathbf{w}, b) = \{\mathbf{z} | \mathbf{w}^\top \mathbf{z} < b, \mathbf{w}, \mathbf{z} \in \mathbb{R}^n, b \in \mathbb{R}\}$ *be a given half space, with* $\mathbf{w} \neq 0$. *Then*

$$P(\mathbf{X} \in \mathcal{H}) \geq \frac{s^2}{s^2 + \mathbf{w}^\top \Sigma \mathbf{w}} \tag{6}$$

where $s = (b - \mathbf{w}^\top \mu)_+$, $(x)_+ = max(x, 0)$.

For proof see Appendix A.

The theorem says that the probability of the event that an observation drawn from $\mathbf{X}$, takes values in the halfspace $\mathcal{H}$ can be upperbounded using the μ and Σ. Let $\mathbf{X}_1 \sim (\mu_1, \Sigma_1)$ denote a class of distributions that have mean μ_1, and covariance Σ_1, but are otherwise arbitrary; likewise for class 2, $\mathbf{X}_2 \sim (\mu_2, \Sigma_2)$. The discriminating hyperplane tries to place class 1 in the half space $\mathcal{H}_1(\mathbf{w}, b) = \{\mathbf{x} | \mathbf{w}^T \mathbf{x} > b\}$ and class 2 in the other half space, $\mathcal{H}_2(\mathbf{w}, b) = \{\mathbf{x} | \mathbf{w}^T \mathbf{x} < b\}$. To ensure this one has to find $\{\mathbf{w}, b\}$ such that $P(\mathbf{X}_1 \in \mathcal{H}_1)$ and $P(\mathbf{X}_2 \in \mathcal{H}_2)$ are both high. Lanckriet et al. (2002a), (also see Lanckriet et al., 2002b), considers this problem and solves it in a minimax setting.

In this paper we consider the problem of feature selection. As remarked before feature selection can be enforced by minimizing the L_1 norm of $\mathbf{w}$. To this end consider the following problem

$$\begin{array}{ll} \underset{\mathbf{w},b}{\min} & \|\mathbf{w}\|_1 \\ s.t & Prob(\mathbf{X}_1 \in \mathcal{H}_1) \geq \eta \\ & Prob(\mathbf{X}_2 \in \mathcal{H}_2) \geq \eta \\ & \mathbf{X}_1 \sim (\mu_1, \Sigma_1) \; \mathbf{X}_2 \sim (\mu_2, \Sigma_2) \end{array} \tag{7}$$

The sparseness criterion is approximately enforced by the objective. In most cases the objective yields a sparse $\mathbf{w}$. The two constraints state that the probability of belonging to the proper half space, should be atleast more than a user defined parameter η. The parameter η takes values in $(0, 1)$. Higher the value of η more stringent is the requirement that all points belong to the correct half space.

The problem (7) has two constraints, one for each class, which states that the lower bound of probability of a random vector taking values in a given half space is η. These constraints can be posed as nonlinear constraints by applying theorem 1 (see Lanckriet et al., 2002b). The constraint for class 1 can be handled by setting

$$Prob(\mathbf{x}_1 \in \mathcal{H}_1) \geq \frac{(\mathbf{w}^T\mu_1 - b)_+^2}{(\mathbf{w}^T\mu_1 - b)_+^2 + \mathbf{w}^T\Sigma\mathbf{w}} \geq \eta$$

which yield two constraints

$$\mathbf{w}^\top\mu_1 - b \geq \sqrt{\frac{\eta}{1-\eta}}\sqrt{\mathbf{w}^T\Sigma_1\mathbf{w}} \quad \mathbf{w}^\top\mu_1 - b \geq 0$$

Similarly applying theorem 1 to the other constraint, two more constraints are obtained. Note that the constraints are positively homogenous, that is if $\mathbf{w}, b$ satisfies the constraints then a $c\mathbf{w}, cb$ also satisfies the constraints; c is a positive number. To deal with this extra degree of freedom, one can impose the constraint that the classifier should separate μ_1 and μ_2 even if $\eta = 0$. One way to impose this is via the constraint

$$\mathbf{w}^T\mu_1 - b \geq 1 \quad b - \mathbf{w}^T\mu_2 \geq 1$$

Since both the matrices Σ_1 and Σ_2 are positive semi-definite, there exist matrices $\mathbf{C}_1$ and $\mathbf{C}_2$ such that

$$\Sigma_1 = \mathbf{C}_1\mathbf{C}_1^\top \quad \Sigma_2 = \mathbf{C}_2\mathbf{C}_2^\top$$

The problem (7) can now be stated as a deterministic optimization problem

$$\begin{aligned}
\min_{\mathbf{w},b} \quad & \|\mathbf{w}\|_1 \\
s.t \quad & \mathbf{w}^\top\mu_1 - b \geq \sqrt{\tfrac{\eta}{1-\eta}}\|\mathbf{C}_1^\top\mathbf{w}\|_2 \\
& b - \mathbf{w}^\top\mu_2 \geq \sqrt{\tfrac{\eta}{1-\eta}}\|\mathbf{C}_2^\top\mathbf{w}\|_2 \\
& \mathbf{w}^T\mu_1 - b \geq 1 \\
& b - \mathbf{w}^T\mu_2 \geq 1
\end{aligned}$$

The nonlinear objective can be tackled, as in (4), by introducing two vectors $\mathbf{u}$ and $\mathbf{v}$, and which leads to the formulation

$$\begin{aligned}
\min_{\mathbf{u},\mathbf{v},b} \quad & (\mathbf{u}+\mathbf{v})^\top \mathbf{e} \\
s.t \quad & (\mathbf{u}-\mathbf{v})^\top \mu_1 - b \geq \sqrt{\tfrac{\eta}{1-\eta}} \|\mathbf{C}_1^\top (\mathbf{u}-\mathbf{v})\|_2 \\
& b - (\mathbf{u}-\mathbf{v})^\top \mu_2 \geq \sqrt{\tfrac{\eta}{1-\eta}} \|\mathbf{C}_2^\top (\mathbf{u}-\mathbf{v})\|_2 \\
& (\mathbf{u}-\mathbf{v})^\top \mu_1 - b \geq 1 \\
& b - (\mathbf{u}-\mathbf{v})^\top \mu_2 \geq 1 \\
& \mathbf{u} \geq 0, \mathbf{v} \geq 0
\end{aligned} \tag{8}$$

This problem is convex, and is an instance of SOCP. The nonlinear constraints are called Second Order Cone(SOC) constraints. A SOC constraint on the variable $\mathbf{x} \in \mathbb{R}^n$ is of the form

$$\mathbf{c}^\top \mathbf{x} + d \geq \|\mathbf{A}\mathbf{x} + \mathbf{b}\|_2$$

where $\mathbf{b} \in \mathbb{R}^m, \mathbf{c} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{m \times n}$ are given. Linear constraints are special case of such constraints. Minimizing a linear objective over SOC constraints is known as SOCP problems. These problems can be solved efficiently by publicly available codes: recent advances in Interior point methods for convex nonlinear optimization (Nesterov and Nemirovskii, 1993) have made such problems feasible. As a special case of convex nonlinear optimization SOCPs have gained much attention in recent times. For a discussion of further efficient algorithms and applications of SOCP (see Lobo et al., 1998).

On the training dataset the error rate of the classifier is upper bounded by $1-\eta$. The upper bound also holds for the generalization error (Lanckriet et al., 2002b) if the test data comes from a distribution having the same mean and covariance as the training data. As η is increased, the data is forced to lie on the correct side of the hyperplane with more probability. This should result in a smaller training error. Again with increasing η, sparseness would decrease, as more stress is given to accuracy. Thus the parameter η trades off accuracy with sparseness.

3.1 Feature selection for Fisher discriminants

In this section we derive the fisher discriminant using the chebychev bound and discuss a formulation for feature selection. For a linearly separable dataset all observations belonging to class 1(class 2) obeys $\mathbf{w}^\top \mathbf{x}_1 \geq b$ ($\mathbf{w}^\top \mathbf{x}_2 \leq b$) which implies that $\mathbf{w}^\top \mathbf{X}_1 \geq b (\mathbf{w}^\top \mathbf{X}_2 \leq b)$ for a carefully chosen $\{\mathbf{w}, b\}$. If $\mathbf{X} = \mathbf{X}_1 - \mathbf{X}_2$ defines the difference between the class conditional random vectors, then $\mathbf{X}$ lies in the halfspace $\mathcal{H}(\mathbf{w}) = \{\mathbf{z} | \mathbf{w}^\top \mathbf{z} \geq 0\}$. One can derive the fisher discriminant by considering the following formulation

$$\begin{aligned}
\max_{\mathbf{w},\eta} \quad & \eta \\
s.t \quad & Prob(\mathbf{X} \in \mathcal{H}) \geq \eta \quad \mathbf{X} \sim (\mu, \Sigma)
\end{aligned} \tag{9}$$

Since $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent the mean of $\mathbf{X}$ is $\mu = \mu_1 - \mu_2$ and covariance $\Sigma = \Sigma_1 + \Sigma_2$. Using the Chebychev bound (6) the constraint can be lower bounded by

$$Prob(\mathbf{X} \in \mathcal{H}) \geq \frac{(\mathbf{w}^\top \mu)^2}{(\mathbf{w}^\top \mu)^2 + \mathbf{w}^\top \Sigma \mathbf{w}} \quad \mathbf{w}^\top \mu \geq 0$$

and hence it follows that (9) is equivalent to solving

$$\max_{\mathbf{w}} \frac{\{\mathbf{w}^\top(\mu_1 - \mu_2)\}^2}{\mathbf{w}^\top(\Sigma_1 + \Sigma_2)\mathbf{w}} \tag{10}$$

which is same as the fisher discriminant. The above formulation shows that fisher discriminant can be understood as computing a discriminant hyperplane whose generalization error is less than $1 - \eta^*$, where

$$\eta^* = \frac{d(\mathbf{w}^*)}{1 + d(\mathbf{w}^*)} \quad d(\mathbf{w}^*) = \max_{\mathbf{w}} \frac{\{\mathbf{w}^\top(\mu_1 - \mu_2)\}^2}{\mathbf{w}^\top(\Sigma_1 + \Sigma_2)\mathbf{w}}$$

The bound holds provided the data distribution has the neccessary first and second moments. One can incorporate feature selection by minimizing the L_1 norm of $\mathbf{w}$ for a fixed value of η as follows

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \|\mathbf{w}\|_1 \\ s.t \quad & Prob(\mathbf{X} \in \mathcal{H}) \geq \eta \quad \mathbf{X} \sim (\mu_1 - \mu_2, \Sigma_1 + \Sigma_2) \end{aligned} \tag{11}$$

and arguing as in (8) the following SOCP

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \|\mathbf{w}\|_1 \\ s.t \quad & \mathbf{w}^\top(\mu_1 - \mu_2) \geq \sqrt{\tfrac{1-\eta}{\eta}}\sqrt{\mathbf{w}^\top(\Sigma_1 + \Sigma_2)\mathbf{w}} \\ & \mathbf{w}^\top(\mu_1 - \mu_2) \geq 1 \end{aligned} \tag{12}$$

is obtained. The parameter η ensures that the resulting classifier has a misclassification error less than $1 - \eta$, while feature selection is ensured by the objective.

3.2 Estimation of mean and covariance for each class

Let $T_1 = [\mathbf{x}_{11}, \ldots, \mathbf{x}_{1m_1}]$ be the data matrix for one class, say with label $y = 1$. Similarly $T_2 = [\mathbf{x}_{21}, \ldots, \mathbf{x}_{2m_2}]$ be the data matrix for the other class having the label $y = -1$. Both the matrices have the same number of rows n, the number of features. The columns correspond to datapoints; m_1 datapoints for the first class and m_2 datapoints for the other class. For Microarray datasets the number of features, n is in thousands, while the number of examples m_1 or m_2 is less than hundred.

In the present formulation empirical estimates of the mean and covariance are used

$$\mu_1 = \overline{\mathbf{x}}_1 = \frac{1}{m_1} T_1 \mathbf{e} \;\; \mu_2 = \overline{\mathbf{x}}_2 = \frac{1}{m_2} T_2 \mathbf{e}$$

$$\Sigma_1 = \overline{\Sigma}_1 = \mathbf{C}_1\mathbf{C}_1^T \;\; \mathbf{C}_1 = \frac{1}{\sqrt{m_1}}(T_1 - \mu_1\mathbf{e}^\top)$$

$$\Sigma_2 = \overline{\Sigma}_2 = \mathbf{C}_2\mathbf{C}_2^T \;\; \mathbf{C}_2 = \frac{1}{\sqrt{m_2}}(T_2 - \mu_2\mathbf{e}^\top)$$

Note that the covariances are huge matrices (of size $n \times n$), instead one can store the much smaller matrices C_1 and C_2 of size $n \times m_1$ and $n \times m_2$. The resulting classifier is heavily dependent on the estimates of the mean and covariance. In the next section we will discuss classifiers which are robust to errors in the estimation of mean and covariance.

4. A robust formulation

In practical cases it might happen that the error rate of the classifiers is well above $1-\eta$. As pointed out in Lanckriet et al. (2002b), this problem often occurs when the training dataset has very few data-points compared to the number of features, for example microarray datasets. In such cases the estimates of mean and covariance are not very accurate. It would be useful, especially for microarray datasets, to explore formulations which can yield classifiers robust to such estimation errors. In the following we discuss one such formulation.

We assume that the means and covariances take values in a specified set, in particular $(\mu_1, \Sigma_1) \in U_1$ where $U_1 \subset \mathbb{R}^n \times S_n^+$, where S_n^+ is the set of all positive semidefinite $n \times n$ matrices. Similarly another set U_2 is defined which characterizes the values of (μ_2, Σ_2). Consider the robust version of formulation (7),

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \|\mathbf{w}\|_1 \\ s.t \quad & Prob(\mathbf{X}_1 \in \mathcal{H}_1) \geq \eta \\ & Prob(\mathbf{X}_2 \in \mathcal{H}_2) \geq \eta \\ & \mathbf{X}_1 \sim (\mu_1, \Sigma_1)\ \mathbf{X}_2 \sim (\mu_2, \Sigma_2) \\ & (\mu_1, \Sigma_1) \in U_1,\ (\mu_2, \Sigma_2) \in U_2 \end{aligned} \tag{13}$$

It ensures that the misclassification rate of the classifier is always less than $1 - \eta$ for any arbitrary distribution whose means and covariances take values in some specified sets.

The tractability of this formulation depends on the definition of the sets U_1 and U_2. We assume that the sets describing the values of means and covariances are independent of one another, more precisely $U_{m1}, U_{m2}, U_{v1}, U_{v2}$ describe the uncertainty in the values of $\mu_1, \mu_2, \Sigma_1, \Sigma_2$ respectively. As before applying the chebychev bound and with a reformulation of (8) the following robust version

$$\begin{aligned} \min_{\mathbf{w},b,t_1,t_2} \quad & \|\mathbf{w}\|_1 \\ s.t. \quad & \mathbf{w}^\top\mu_1 - b \geq t_1 \ \forall \mu_1 \in U_{m1} \\ & b - \mathbf{w}^\top\mu_2 \geq t_2 \ \forall \mu_2 \in U_{m2} \\ & \sqrt{\mathbf{w}^\top\Sigma_1\mathbf{w}} \leq \sqrt{\tfrac{1-\eta}{\eta}}t_1 \ \forall \Sigma_1 \in U_{v1} \\ & \sqrt{\mathbf{w}^\top\Sigma_2\mathbf{w}} \leq \sqrt{\tfrac{1-\eta}{\eta}}t_2 \ \forall \Sigma_2 \in U_{v2} \\ & t_1 \geq 1 \;\; t_2 \geq 1 \end{aligned} \tag{14}$$

is obtained. The reformulation is obtained by modifying the SOC constraint corresponding to class 1 by introducing a new variable t_1 as follows,

$$\mathbf{w}^\top \mu_1 - b \geq t_1 \geq \sqrt{\frac{\eta}{1-\eta}} \|\mathbf{C}_1^\top \mathbf{w}\|_2 \;\; t_1 \geq 1$$

Likewise another variable is introduced to deal with the other SOC constraint belonging to class 2. To restrict the uncertainty to a low dimension space the following assumption is made.

Assumption 2 *The random vector $\mathbf{X}_1$ take values in the linear span of columns of T_1, while the random vector $\mathbf{X}_2$ take values in the linear span of columns of T_2. More precisely the n dimensional random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ are linearly related to a m_1 dimensional random vector $\mathbf{Z}_1$ and a m_2 dimensional random vector$\mathbf{Z}_2$ respectively as follows*

$$\mathbf{X}_1 = T_1 \mathbf{Z}_1 \qquad \mathbf{X}_2 = T_2 \mathbf{Z}_2 \tag{15}$$

For microarray datasets m_1 and m_2 are much smaller than n. Thus the assumption restricts the random variables $\mathbf{X}_1$ and $\mathbf{X}_2$ to much smaller dimension spaces. Let μ_{z1}, Σ_{z1} be the mean and covariance of the random variable $\mathbf{Z}_1$, and μ_{z2}, Σ_{z2} be the mean and covariance of the random variable $\mathbf{Z}_2$, it follows that

$$\mu_1 = T_1 \mu_{z1} \;\; \mu_2 = T_2 \mu_{z2} \;\; \Sigma_1 = T_1 \Sigma_{z1} T_1^\top \;\; \Sigma_2 = T_2 \Sigma_{z2} T_2^\top$$

Clearly then the sample estimates $\overline{\mathbf{z}}_1$ and $\overline{\mathbf{z}}_2$ are related to the sample estimates $\overline{\mathbf{x}}_1$ and $\overline{\mathbf{x}}_2$ by

$$\overline{\mathbf{x}}_i = T_i \overline{\mathbf{z}}_i \quad \overline{\Sigma}_i = T_i \overline{\Sigma}_{zi} T_i^\top \quad \overline{\Sigma}_{zi} = \frac{1}{m_i}(\mathbf{I} - \mathbf{e}\mathbf{e}^\top) \forall i \in \{1, 2\} \tag{16}$$

Assuming an ellipsoidal uncertainty on the estimate of μ_1 and in light of (16), we define

$$U_{m1} = \{\mu_1 | \mu_1 = T_1 \mu_{z1}, \quad (\mu_{z1} - \overline{\mathbf{z}}_1)^\top T_1^\top T_1 (\mu_{z1} - \overline{\mathbf{z}}_1) \leq \delta^2\}$$

For the uncertainty set U_{m1} and a given $\mathbf{w}$ the constraint

$$\mathbf{w}^\top \mu_1 - b \geq t_1 \; \forall \mu_1 \in U_{m1}$$

is equivalent to

$$\min_{\mu_1 \in U_{m1}} \mathbf{w}^\top \mu_1 - b \geq t_1 \tag{17}$$

Noting that minimizing a linear function over an ellipsoid is a convex optimization problem, which has the following closed form solution (see Appendix B)

$$\min_{\mu_1 \in U_{m1}} \mathbf{w}^\top \mu_1 = \frac{1}{m_1} \mathbf{w}^\top T_1 \mathbf{e} - \delta \|T_1 \mathbf{w}\| \tag{18}$$

the constraint (17) can be restated as

$$\frac{1}{m_1} \mathbf{w}^\top T_1 \mathbf{e} - b \geq t_1 + \delta \|T_1 \mathbf{w}\| \tag{19}$$

Similarly for μ_2 the uncertainty set is defined as

$$U_{m2} = \{\mu_2 | \mu_2 = T_2\mu_{z2}, \quad (\mu_{z2} - \overline{\mathbf{z}}_2)^\top T_2^\top T_2(\mu_{z2} - \overline{\mathbf{z}}_2) \leq \delta^2\}$$

and analogous to (19) the following constraint is

$$b - \frac{1}{m_2}\mathbf{w}^\top T_2\mathbf{e} \geq t_2 + \delta\|T_2\mathbf{w}\| \tag{20}$$

is obtained. Following Lanckriet et al. (2002b), the sets characterizing the covariance matrices are defined using Frobenius norm

$$U_{vi} = \{\Sigma_i | \Sigma_i = T_i\overline{\Sigma}_{zi}T_i^\top \;\; \|\Sigma_{zi} - \overline{\Sigma}_{zi}\| \leq \rho\} \;\; i = \{1, 2\}$$

Imposing robustness to estimation errors in the covariance matrix Σ_i is equivalent to the constraint

$$\max_{\Sigma_i \in U_{vi}} \sqrt{\mathbf{w}^\top \Sigma_i \mathbf{w}} \leq \sqrt{\frac{1-\eta}{\eta}} t_i \;\; i = \{1, 2\} \tag{21}$$

Using the result (see Appendix B)

$$\max_{\Sigma_i \in U_{vi}} \sqrt{\mathbf{w}^\top \Sigma_i \mathbf{w}} = \sqrt{\mathbf{w}^\top T_i(\overline{\Sigma}_{zi} + \rho\mathbf{I})T_i^\top \mathbf{w}}$$

the formulation (14) turns out to be a

$$\begin{aligned} \min_{\mathbf{w},b,t_1,t_2} \quad & \|\mathbf{w}\|_1 \\ s.t. \quad & \tfrac{1}{m_1}\mathbf{w}^\top T_1\mathbf{e} - b \geq t_1 + \delta\|T_1\mathbf{w}\|_2 \\ & b - \tfrac{1}{m_2}\mathbf{w}^\top T_2\mathbf{e} \geq t_2 + \delta\|T_2\mathbf{w}\|_2 \\ & \|\mathbf{C}_{1z}^\top T_1^\top \mathbf{w}\|_2 \leq \sqrt{\tfrac{1-\eta}{\eta}} t_1 \\ & \|\mathbf{C}_{2z}^\top T_2^\top \mathbf{w}\|_2 \leq \sqrt{\tfrac{1-\eta}{\eta}} t_2 \\ & t_1 \geq 1 \;\; t_2 \geq 1 \end{aligned} \tag{22}$$

SOCP. The matrix $\mathbf{C}_{1z}$ is obtained by using the cholesky decomposition of the regularized matrix $\overline{\Sigma}_{z1} + \rho I$, similarly for $\mathbf{C}_{2z}$.

As a consequence of Assumption 2 one needs to factorize matrices of size $m_1 \times m_1$, and $m_2 \times m_2$ instead of a much larger $n \times n$ matrix for the Frobenius norm uncertainty model. Thus the assumption has computational benefits but it is quite restrictive. However in absence of any prior knowledge this maybe a good alternative to explore. In the next section we experiment on the formulations (8), and (22) on both synthetic and real world microarray datasets.

5. Experiments

The feature selection abilities of the proposed formulations were tested on both synthetic and real world datasets. As a benchmark the performance of the LP formulation on the same datasets are also reported.

η	0	0.2	0.4	0.6	0.8	0.9	0.95	0.99
fs	10	10	10	10	10	10	9, 10	7,8,9,10

Table 1: The set of selected features, fs, for various values of η on synthetic dataset. See text for more details

Consider a synthetic dataset generated as follows. The class label,y, of each observation was randomly chosen to be 1 or -1 with probability 0.5. The first ten features of the observation, $\mathbf{x}$, are drawn as $y\mathcal{N}(-i, 1)$, where $\mathcal{N}(\mu, \sigma^2)$ is a gaussian centered around μ and with variance σ^2. Nine hundred ninety other features were drawn as $\mathcal{N}(0, 1)$. Fifty such observations were generated. The feature selection problem is to detect the first ten features, since they are the most discriminatory from the given pool of 1000 features, when the sample size is fifty.

The features were selected by the following procedure (see Murray, 1977). From the dataset fifty partitions was generated by holding out one example as test data and others as training data. For each partiton formulation (8) was solved on the training data, for a fixed value of η, using the open source package SEDUMI Sturm (1999) to obtain a set of features, and the resulting classifier was used to predict the label of test data. The union of fifty sets of features is reported in Table 1 for various values of η. The average number of errors on all the fifty test sets, the Leave one out(LOO) error, was found to be 0. For low values of η say $\eta = 0.2$, only one feature, feature number 10, was selected. This is not surprising because among the ten features, feature number 10 has the most discriminatory power As the value of η is increased the formulation reports more discriminatory features. For $\eta = 0.95$, four features feature numbers 7,8,9,10, was selected. This shows that, inspite of sample size being low compared to the number of features this formulation is able to discover the most discriminatory features. The corresponding list of features selected by the linear programming formulation (5), are fs $= \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Both the formulations pick up the most discriminatory features. The experiment was repeated for 100 randomly generated datasets, and gave similar results. This demonstrates that the formulation (8) picks up discriminatory features and is comparable to the LP formulation.

We also experimented with the robust formulation (14) for different values of δ, and ρ. In figure (1) number of features are plotted for various values of δ, As δ increases the number of features selected by the formulation (22) increases, the value of ρ was zero for the reported experiment. The robust formulation tries to maintain the classification accuracy even when the estimates of mean and covariance are not correct. To ensure this more and more features are needed to maintain the accuracy. Similar results were also obtained by varying ρ.

The formulation (8) was tested on six datasets, B,C,D,E,F,G defined in (Bhattacharyya et al., 2003). These binary classification problems are related to Small Round Blue Cell Tumors (SRBCT). Each dataset have various number of data points, but have the same number of features, $n = 2308$.

From the given dataset a partition was generated by holding out a datapoint as test set while the training set consisted of all the other datapoints. For each fixed value of η the

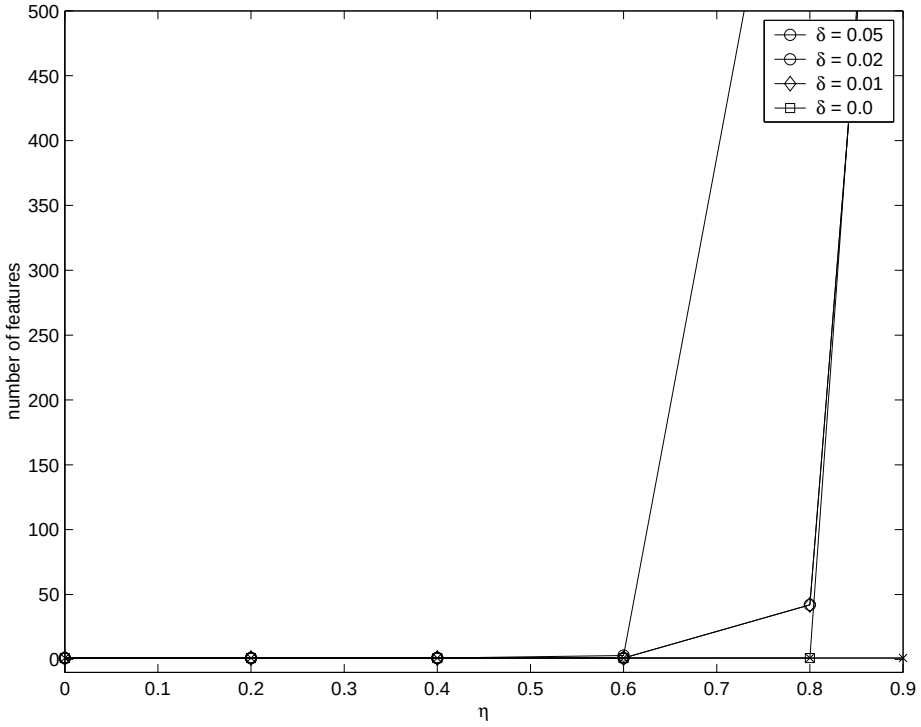

Figure 1: Plots of number of selected features versus η for various values of δ

dataset	B	C	D	E	F	G
SOCP	38	46	25	1	23	21
LP	21	8	8	2	12	13

Table 2: Number of selected features for which the LOO error was minimum. The row titled SOCP tabulates the minimum number of features reported by (8) for which the LOO error was zero. The row titled LP tabulates the number of features selected by the LP formulation. Total number of features is $n = 2308$

formulation was solved for all possible partitions. For each partition the resulting classifier was tested on the held out datapoint. Average number of errors over all the partitions was reported as the LOO (leave one out) error. The results were compared against the linear programming formulation(5).

Table 2 compares the number of features required to attain a zero LOO error by formulations (8) and (5). In both cases a very small number of features, less than 2% of the total number of 2308 features, was selected. However the LP formulations almost always found a smaller set of features. Figures (2, 3, 4) show plots for the number of features selected by (8) for various values of η. As η increases the number of features increase. For comparison the number of features selected by the LP formulation is also plotted on the same graph.

Figures (5,6, 7) show plots for the LOO error the SOCP formulation. As η increases the LOO error decreases. This conforms to the view that as η increases the classifier is forced to be accurate which leads to increase in the number of features.

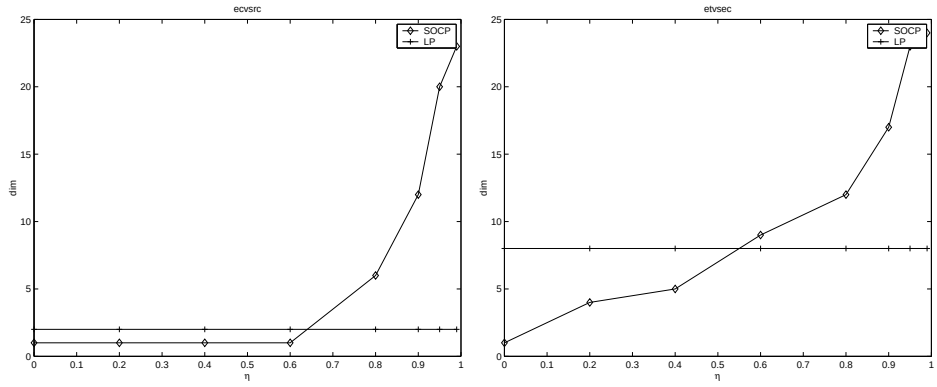

Figure 2: Plots of number of selected features versus η for datasets E, F

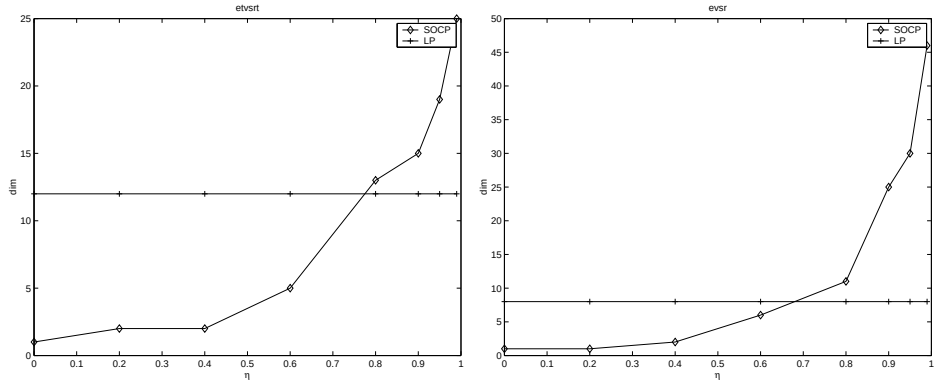

Figure 3: Plots of number of selected features versus η for datasets D, C

6. Conclusions and Future Directions

The problem of selecting discriminatory features by using the moments of the class conditional densities was addressed in the paper. Using a Chebyshev-Cantelli inequality, the problem was posed as a SOCP. The above approach was also used to derive a formulation for doing feature selection for fisher discriminants. A robust formulation was discussed which yields classifiers robust to estimation errors in the mean and covariance.

On a toy dataset the formulation discovered the discriminatory features. The formulation has a parameter η, which can trade off accuracy with number of features. For small values of η low number of discriminatory features are reported, while as η is increased

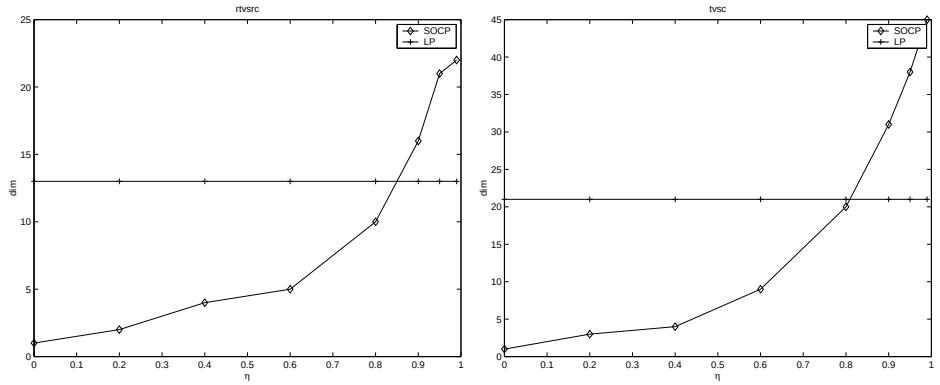

Figure 4: Plots of number of selected features versus η for datasets G, B.

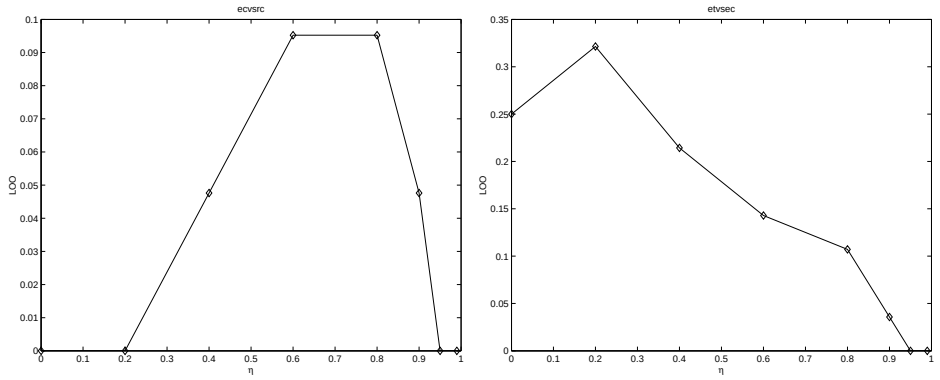

Figure 5: Plots of LOO error versus η for datasets E, F

the formulation reports more number of features. As borne out by experiments on the microarray datasets, the accuracy of the classifier increases as η is increased.

The approach in this paper can also be extended to design nonlinear classifiers using very few support vectors. Let the discriminating surface be $\{\mathbf{x}|\sum_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) = b\}$ which divides the n-dimensional Euclidean space into two disjoint subsets $\{\mathbf{x}|\sum_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) < b\}$ and $\{\mathbf{x}|\sum_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) > b\}$, where the kernel K, is a function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ obeying the Mercer conditions (Mercer, 1909).

One can restate the nonlinear discriminating surface by a hyperplane in m dimensions,

$$\mathcal{H} = \{\mathbf{x}|\alpha^\top k(x) - b = 0\}$$

where m is the number of examples and $k(\mathbf{x})$ is a vector in m dimensions whose ith component is $k(\mathbf{x}, \mathbf{x}_i)$. The set of support vectors is defined by $S = \{i|\alpha_i \neq 0\}$. We wish to find a decision surface utilizing very small number of these vectors, or in other words the goal is to minimize the cardinality of the set S, which can beapproximated by the L_1 norm of α.

Let $k_1 = k(\mathbf{X}_1)$ be a random vector corresponding to class 1 while $k_2 = k(\mathbf{X}_2)$ be another random vector belonging to class 2. Let the means of k_1 and k_2 be $\tilde{k}_1$ and $\tilde{k}_2$

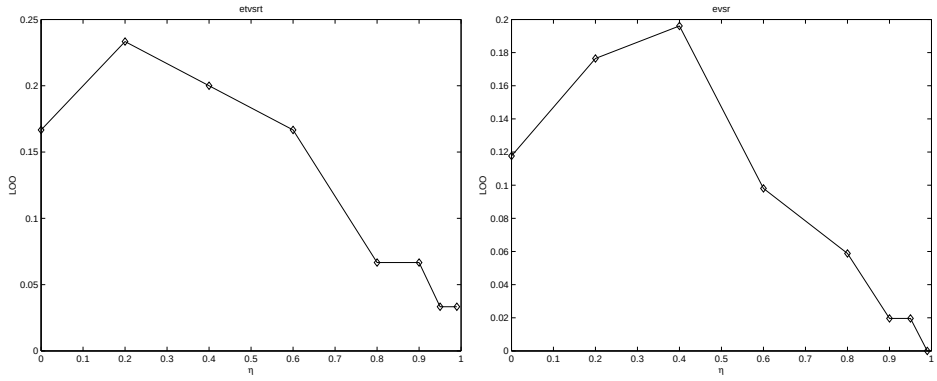

Figure 6: Plots of LOO error versus η for datasets D, C

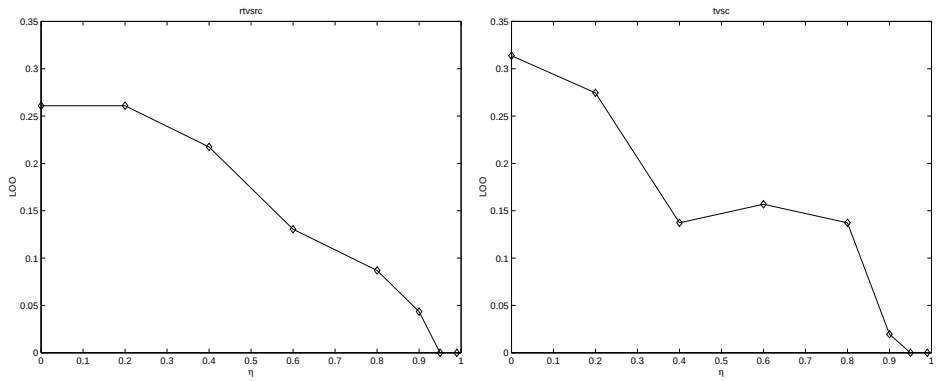

Figure 7: Plots of LOO error versus η for datasets G, B

respectively and the covariance be $\tilde{\Sigma}_1$ and $\tilde{\Sigma}_2$ respectively. The problem can be approached as in (7) and on applying the chebychev bound (6) the following formulation

$$\begin{aligned} \min_{\alpha,b} \quad & \|\alpha\|_1 \\ s.t \quad & \alpha^\top \tilde{k}_1 - b \geq \sqrt{\tfrac{\eta}{1-\eta}} \sqrt{\alpha^\top \tilde{\Sigma}_1^\top \alpha} \\ & b - \alpha^\top \tilde{k}_2 \geq \sqrt{\tfrac{\eta}{1-\eta}} \sqrt{\alpha^\top \tilde{\Sigma}_2^\top \alpha} \\ & \alpha^\top \tilde{k}_1 - b \geq 1 \\ & b - \alpha^\top \tilde{k}_2 \geq 1 \end{aligned} \tag{23}$$

is obtained. We believe this can have non-trivial advantages for data-mining problems.

Acknowledgements

The author thanks the referees and the editor for their constructive comments. Their suggestions improved the paper significantly.

References

E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 1998.

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

S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *Siam Journal of Scientific Computing*, 20(1):33–61, 1999.

R. Fletcher. *Practical Methods of Optimization*. John Wiley and Sons, New York, 1989.

G. R.G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. I. Jordan. Minimax probability machine. In *Advances in Neural Information Processing Systems*. MIT Press, 2002a.

G. R.G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, pages 555 – 582, December 2002b.

M.S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284(1–3):193–228, 1998.

Gábor Lugosi. Concentration-of-measure inequalities, 2003. URL `http://www.econ.upf.es/~lugosi/pre.html/anu.ps`. Lecture notes presented at the Machine learning Summer School 2003, ANU, Canberra.

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

J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London*, A 209: 415–446, 1909.

G.D. Murray. A cautionary note on selection of variables in discriminant analysis. *Applied Statistics*, 26(3):246–250, 1977.

Y. Nesterov and A. Nemirovskii. *Interior Point Algorithms in Convex Programming*. Number 13 in Studies in Applied Mathematics. SIAM, Philadelphia, 1993.

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

J. F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11/12(1-4):625–653, 1999.

J. Weston, A. Elisseff, B. Schölkopf, and M. Tipping. Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Research*, 3, 2003.

Appendix A: The Chebychev-Cantelli inequality

In this appendix we prove a multivariate generalization of one sided chebychev inequality, This inequality will be used to derive a lower bound on the probability of a multivariate random variable taking values in a given half space. Marshall and Olkin (1960) proved a more general case, also see (Popescu and Bertsimas, 2001).

We first state and prove the chebychev inequality for a univariate random variable, see (Lugosi, 2003). It would be useful to recall the Cauchy-Schwartz inequality. Let X and Y be random variables with finite variances, $\mathbb{E}(X - \mathbb{E}(X))^2 < \infty$ and $\mathbb{E}(Y - \mathbb{E}(Y))^2 < \infty$. Then

$$|\mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]| \leq \sqrt{\mathbb{E}(X - \mathbb{E}(X))^2 \mathbb{E}(Y - \mathbb{E}(Y))^2}$$

Chebychev-Cantelli inequality *Let $s \geq 0$, Then*

$$P(X - \mathbb{E}(X) < s) \geq \frac{s^2}{s^2 + \mathbb{E}(X - \mathbb{E}(X))^2}$$

Proof Let $Y = X - \mathbb{E}(X)$. Note that $\mathbb{E}(Y) = 0$. For any s the following is true

$$s = \mathbb{E}(s - Y) \leq \mathbb{E}((s - Y) I_{\{Y < s\}}(Y))$$

For any $s \geq 0$, using the Cauchy-Schwartz inequality

$$\begin{aligned} s^2 \leq & \quad \mathbb{E}(s - Y)^2 \mathbb{E}(I^2_{\{Y<s\}}(Y)) \\ = & \quad \mathbb{E}(s - Y)^2 P(Y < s) \\ = & \quad (\mathbb{E}(Y^2) + s^2) P(Y < s) \end{aligned} \tag{24}$$

On rearranging terms one obtains

$$P(Y < s) \geq \frac{s^2}{\mathbb{E}(Y^2) + s^2}$$

and the result follows. ■

The above inequality can be used to derive a lower bound on the probability of a random vector taking values in a given half space.

Theorem 1 *Let $\mathbf{X}$ be a n dimensional random vector. The mean and covariance of $\mathbf{X}$ be $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$. Let $\mathcal{H}(\mathbf{w}, b) = \{\mathbf{z} | \mathbf{w}^\top \mathbf{z} < b, \mathbf{w}, \mathbf{z} \in \mathbb{R}^n, b \in \mathbb{R}\}$ be a given half space, with $\mathbf{w} \neq 0$. Then*

$$P(\mathbf{X} \in \mathcal{H}) \geq \frac{s^2}{s^2 + \mathbf{w}^\top \Sigma \mathbf{w}}$$

where $s = (b - \mathbf{w}^\top \mu)_+$, $(x)_+ = max(x, 0)$.

Proof There are two cases $b \leq \mathbf{w}^\top \mu$ and $b > \mathbf{w}^\top \mu$.

Consider the case $b \leq \mathbf{w}^T \mu$. For this case $s = 0$, and plugging its value in the Chebychev-Cantelli inequality, yields $P(\mathbf{X} \in \mathcal{H}) \geq 0$ which is trivially true.

Consider the other case $b > \mathbf{w}^T\mu$, by definition $s = b - \mathbf{w}^T\mu$. Define $Y = \mathbf{w}^\top \mathbf{x}$, so that $\mathbb{E}(Y) = \mathbf{w}^T\mu \quad \mathbb{E}(Y - \mathbb{E}(Y))^2 = \mathbf{w}^\top \Sigma \mathbf{w}$. We have

$$P(\mathbf{X} \in \mathcal{H}) = P(Y < b) = P(Y - \mathbb{E}(Y) < s)$$

Application of Chebyshev-Cantelli inequality to the above relationship gives our desired result. This completes the proof. ■

Note that the proof does not require Σ to be invertible. For a more general proof pertaining to convex sets and tightness of the bound see (Marshall and Olkin, 1960, Popescu and Bertsimas, 2001).

Appendix B: Uncertainty in Covariance Matrices

Consider the following problem

$$\begin{array}{rl} \max_\Sigma & \sqrt{\mathbf{w}^\top \Sigma \mathbf{w}} \\ \Sigma & = T\Sigma_z T^\top \\ \|\Sigma_z - \overline{\Sigma}_z\|_F & \leq \rho \end{array} \tag{25}$$

Eliminating the equality constraint the objective can be stated as

$$\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}} = \sqrt{\mathbf{w}^\top T \Sigma_z T^\top \mathbf{w}}$$

Introduce a new variable $\Delta\Sigma \in S_n^+$ such that

$$\Sigma_z = \overline{\Sigma}_z + \Delta\Sigma$$

the optimization problem (25) can be stated as

$$\begin{array}{rl} \max_{\Delta\Sigma} & \sqrt{\mathbf{w}^\top T(\overline{\Sigma}_z + \Delta\Sigma)T^\top \mathbf{w}} \\ s.t. & \|\Delta\Sigma\|_F \leq \rho \end{array} \tag{26}$$

The optimal is achieved at $\Delta\Sigma = \rho I$, see Appendix C in Lanckriet et al. (2002b) for a proof.

Notation

The vector $[1, 1, \ldots, 1]^\top$ will denoted by $\mathbf{e}$, and $[0, \ldots, 0]$ by 0, the dimension will be clear from the context. If $\mathbf{w} = [w_1, \ldots, w_n]^\top$, we write $\mathbf{w} \geq 0$ to mean $w_i \geq 0, \forall i \in \{1, \ldots, n\}$. The euclidean norm of a vector $\mathbf{x} = [x_1, \ldots, x_n]^\top$, will be denoted by $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$, while the 1-norm of $\mathbf{x}$ will be denoted by $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$. The indicator function defined on the set A, denoted by $I_A(x)$, is

$$I_A(x) = \begin{cases} 1 & x \in A \\ 0 & otherwise \end{cases}$$

The cardinality of set A is given by $|A|$.