Robust Novelty Detection with Single Class MPM

Gert R.G. Lanckriet[*]
EECS, U.C. Berkeley
gert@eecs.berkeley.edu

Laurent El Ghaoui
EECS, U.C. Berkeley
elghaoui@eecs.berkeley.edu

Michael I. Jordan
Computer Science and
Statistics, U.C. Berkeley
jordan@cs.berkeley.edu

Abstract

The minimax probability machine (MPM) considers a binary classification problem, where mean and covariance matrix of each class are assumed to be known. Without making any further distributional assumptions, the MPM minimizes the worst-case probability $1-\alpha$ of misclassification of future data points. However, the validity of the upper bound $1-\alpha$ depends on the accuracy of the estimates of the real but unknown means and covariances. First, we show how to make this minimax approach robust against certain estimation errors: for unknown but bounded means and covariance matrices, we guarantee a robust upper bound. Secondly, the robust minimax approach for supervised learning is extended in a very natural way to the unsupervised learning problem of quantile estimation – computing a minimal region in input space where at least a fraction α of the total probability mass lives. Mercer kernels can be exploited in this setting to obtain nonlinear regions. Positive empirical results are obtained when comparing this approach to single class SVM and a 2-class SVM approach.

1 Introduction

The minimax probability machine (MPM) considers the supervised learning problem of binary classification. The mean and covariance matrix of each class are assumed to be known, but no further distributional assumptions are made with respect to the class-conditional densities. Under all possible choices of class-conditional densities with given mean and covariance matrix, the MPM mimimizes the worst-case (maximum) probability of misclassification of future data points. This essentially involves exploiting the following powerful theorem presented in [3], for $\alpha \in [0,\ 1)$:

$$\inf_{\mathbf{y}\sim(\bar{\mathbf{y}},\boldsymbol{\Sigma}_\mathbf{y})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \alpha \Leftrightarrow b - \mathbf{a}^T\bar{\mathbf{y}} \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_\mathbf{y}\mathbf{a}},\ \kappa(\alpha) = \sqrt{\alpha/1-\alpha}. \quad (1)$$

where $\bar{\mathbf{y}}, \Sigma_y$ are assumed to be known and $\mathbf{a} \neq 0, b$ given with $\mathbf{a}^T\bar{\mathbf{y}} \leq b$; $\mathbf{y}$ is a random vector and the infimum is taken over all distributions for $\mathbf{y}$ having mean $\bar{\mathbf{y}}$ and covariance matrix $\boldsymbol{\Sigma}_\mathbf{y}$ (assumed to be positive definite for simplicity).

[*] http://robotics.eecs.berkeley.edu/˜gert/

In practice however, $\bar{\mathbf{y}}, \Sigma_y$ are usually unknown *a priori* and need to be estimated from the data. The validity of (1) depends of course on how good the estimates are. After revisiting the MPM shortly, we will show in Section 2 how robust optimization can be used to generalize (1) to give a worst-case guarantee if the mean and covariance are unknown but bounded within some convex region.

Section 3 will then use the robust version of (1) to extend the minimax approach for supervised learning in a very natural way to unsupervised learning: single class MPM. Most general, unsupervised learning involves estimation of the density from which given data points $\mathbf{y}$ are drawn. A simplified version of this problem estimates quantiles of that distribution [1],[4]: for $\alpha \in (0,1]$, one computes a region $\mathcal{Q}$ such that $\mathbf{Pr}\{\mathbf{y} \in \mathcal{Q}\} = \alpha$. For α close to 1, this amounts to novelty detection [4],[5]: the probability density will mostly live in $\mathcal{Q}$ and a data point outside $\mathcal{Q}$ can be considered as significantly different from the given data set. Nonlinear decision boundaries can be obtained using Mercer kernels. Empirical results are given in Section 4.

XXX Mike, I didn't state anything yet in this introduction about the importance of the single class problem... maybe you can add a short and accurate sketch, if necessary/possible? XXX

2 Robust Minimax Probability Machine (R-MPM)

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ denote random vectors in a binary classification problem, modelling data from each of two classes, with means and covariance matrices given by $\bar{\mathbf{x}}, \bar{\mathbf{y}} \in \mathbb{R}^n$, and $\mathbf{\Sigma_x}, \mathbf{\Sigma_y} \in \mathbb{R}^{n \times n}$ (both symmetric and positive semidefinite), respectively. 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 $b \in \mathbb{R}$, that maximizes the worst-case probability α that future data points are classified correctly with respect to all distributions having these means and covariance matrices:

$$\max_{\alpha, \mathbf{a} \neq 0, b} \ \alpha \quad \text{s.t.} \quad \inf_{\mathbf{x} \sim (\bar{\mathbf{x}}, \mathbf{\Sigma_x})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x} \geq b\} \geq \alpha \qquad (2)$$

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

where $\mathbf{x} \sim (\bar{\mathbf{x}}, \mathbf{\Sigma_x})$ refers to the class of distributions that have mean $\bar{\mathbf{x}}$ and covariance $\mathbf{\Sigma_x}$, but are otherwise arbitrary; likewise for $\mathbf{y}$. The worst-case probability of misclassification is explicitly obtained and given by $1 - \alpha$.

The core result in solving the previous optimization problem (2) involves expressing the probabilistic constraints in a deterministic way and is given in (1). This finally leads to the following convex optimization problem [2], to be solved to determine an optimal hyperplane $\mathcal{H}(\mathbf{a}, b)$ [3]:

$$\kappa_*^{-1} := \min_{\mathbf{a}} \sqrt{\mathbf{a}^T\mathbf{\Sigma_x}\mathbf{a}} + \sqrt{\mathbf{a}^T\mathbf{\Sigma_y}\mathbf{a}} \quad \text{s.t.} \quad \mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}}) = 1, \qquad (3)$$

where b is set to the value $b_* = \mathbf{a}_*^T\bar{\mathbf{x}} - \kappa_*\sqrt{\mathbf{a}_*^T\mathbf{\Sigma_x}\mathbf{a}_*}$, with $\mathbf{a}_*$ an optimal solution of (3). The optimal worst-case misclassification probability is obtained via $1 - \alpha_* = 1/1 + \kappa_*^2$. Once an optimal hyperplane is found, classification of a new data point $\mathbf{z}_{new}$ is done by evaluating $\text{sign}(\mathbf{a}_*^T\mathbf{z}_{new} - b_*)$: 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 $\mathbf{y}$.

In practical experiments, it may well happen that the test set error is greater than $1 - \alpha_*$. This contradicts the previous claim that $1 - \alpha_*$ is an upper bound on misclassification error. It seems to imply that the probabilistic guarantees given by the constraints in (2) do not hold. 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 constraints in (2) and thus of the upper bound $1-\alpha^*$ depends on how good the estimates are.

We can approach this estimation problem based on robust optimization. Assume the mean and covariance matrix of each class are unknown but bounded within some specified convex sets: $(\bar{\mathbf{x}}, \mathbf{\Sigma_x}) \in \mathcal{X}$, with $\mathcal{X} \subset \mathbb{R}^n \times \{M \in \mathbb{R}^{n\times n} | M = M^T, M \succeq 0\}$. Likewise $\mathcal{Y}$ describes uncertainty in the mean and covariance matrix of the random variable $\mathbf{y}$. We now want the probabilistic guarantees in (2) to be robust against variations of the mean and covariance matrix within this uncertainty sets:

$$\begin{aligned}\max_{\alpha, \mathbf{a}\neq 0, b} \ \alpha \quad \text{s.t.} \quad & \inf_{\mathbf{x}\sim(\bar{\mathbf{x}},\mathbf{\Sigma_x})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x} \geq b\} \geq \alpha \ \ \forall\, (\bar{\mathbf{x}}, \mathbf{\Sigma_x}) \in \mathcal{X}, \\ & \inf_{\mathbf{x}\sim(\bar{\mathbf{y}},\mathbf{\Sigma_y})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \alpha \ \ \forall\, (\bar{\mathbf{y}}, \mathbf{\Sigma_y}) \in \mathcal{Y}.\end{aligned} \tag{4}$$

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. The complexity of this problem depends obviously on the structure of the uncertainty sets $\mathcal{X}, \mathcal{Y}$. We now consider a specific choice for $\mathcal{X}$ and $\mathcal{Y}$, both realistic from a statistical viewpoint and tractable numerically:

$$\begin{array}{rcl}\mathcal{X} & = & \left\{(\bar{\mathbf{x}}, \mathbf{\Sigma_x}) \ : \ (\bar{\mathbf{x}} - \bar{\mathbf{x}}^0)^T \mathbf{\Sigma_x}^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{x}}^0) \leq \nu^2, \ \ \|\mathbf{\Sigma_x} - \mathbf{\Sigma_x}^0\|_F \leq \rho\right\}, \\ \mathcal{Y} & = & \left\{(\bar{\mathbf{y}}, \mathbf{\Sigma_y}) \ : \ (\bar{\mathbf{y}} - \bar{\mathbf{y}}^0)^T \mathbf{\Sigma_y}^{-1} (\bar{\mathbf{y}} - \bar{\mathbf{y}}^0) \leq \nu^2, \ \ \|\mathbf{\Sigma_y} - \mathbf{\Sigma_y}^0\|_F \leq \rho\right\},\end{array} \tag{5}$$

with $\bar{\mathbf{x}}^0, \mathbf{\Sigma_x}^0$ the "nominal" mean and covariance estimates and with $\nu, \rho \geq 0$ fixed and, for simplicity, assumed equal for $\mathcal{X}$ and $\mathcal{Y}$). Section 4 discusses how their values can be determined. The matrix norm is the Frobenius norm: $\|A\|_F^2 = \mathbf{Tr}(A^T A)$.

Our model for the mean uncertainty assumes the mean of class $\mathbf{y}$ belongs to an ellipsoid centered around $\bar{\mathbf{y}}^0$, with shape determined by the (unknown) $\mathbf{\Sigma_y}$. This is motivated by the standard statistical approach to estimating a region of confidence based on Laplace (i.e., second-order) approximations to a likelihood function. The covariance matrix belongs to a matrix norm ball centered around $\mathbf{\Sigma_y}^0$. This uncertainty model is perhaps less classical from a statistical viewpoint, but it will lead to a regularization term of a classical form.

In order to solve problem (4), we need a robust version of the core result (**??**). Notice that

$$\inf_{\mathbf{y}\sim(\bar{\mathbf{y}},\mathbf{\Sigma_y})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \alpha \quad \forall (\bar{\mathbf{y}}, \mathbf{\Sigma_y}) \in \mathcal{Y} \Leftrightarrow b - \mathbf{a}^T\bar{\mathbf{y}} \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\mathbf{\Sigma_y}\mathbf{a}} \quad \text{with} \quad \kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}} \forall (\bar{\mathbf{y}}, \mathbf{\Sigma_y}) \in \mathcal{Y} \tag{6}$$

where the last part guarantees the constraint holds for the worst-case possible estimate/choice of the mean and covariance matrix within the bounded set $\mathcal{Y}$.

For a given $\mathbf{a}$ and $\bar{\mathbf{x}}^0$, we have

$$\min_{\bar{\mathbf{x}} \ : \ (\bar{\mathbf{x}} - \bar{\mathbf{x}}^0)^T \mathbf{\Sigma_x}^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{x}}^0) \leq \nu^2} \mathbf{a}^T\bar{\mathbf{x}} = \mathbf{a}^T\bar{\mathbf{x}}^0 - \nu\sqrt{\mathbf{a}^T\mathbf{\Sigma_x}\mathbf{a}}. \tag{7}$$

Indeed, we can form the Lagrangian

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

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

At the optimum:

$$\begin{aligned}
\frac{\partial}{\partial \bar{\mathbf{x}}}\mathcal{L}(\bar{\mathbf{x}},\lambda) &= \mathbf{a} + 2\lambda {\boldsymbol{\Sigma}_{\mathbf{x}}}^{-1}\bar{\mathbf{x}} - 2\lambda {\boldsymbol{\Sigma}_{\mathbf{x}}}^{-1}\bar{\mathbf{x}}^0 = 0 \\
\Rightarrow \bar{\mathbf{x}} &= \bar{\mathbf{x}}^0 - \frac{1}{2\lambda}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{a}, & (8) \\
\frac{\partial}{\partial \lambda}\mathcal{L}(\bar{\mathbf{x}},\lambda) &= (\bar{\mathbf{x}} - \bar{\mathbf{x}}^0)^T {\boldsymbol{\Sigma}_{\mathbf{x}}}^{-1}(\bar{\mathbf{x}} - \bar{\mathbf{x}}^0) - \nu^2 \\
&= \frac{1}{4\lambda^2}\mathbf{a}^T\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{a} - \nu^2 = 0 & (9) \\
\Rightarrow \lambda &= \sqrt{\frac{\mathbf{a}^T\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{a}}{4\nu}}, & (10)
\end{aligned}$$

where (8) is substituted to obtain (9). Substitution of (10) in (8) gives the optimal value for $\bar{\mathbf{x}}$ and leads to (7).

For given $\mathbf{a}$ and $\boldsymbol{\Sigma}^0$, we have

$$\max_{\boldsymbol{\Sigma}\,:\,\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^0\|_F\leq\rho} \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} = \mathbf{a}^T\left(\boldsymbol{\Sigma}^0 + \rho I_n\right)\mathbf{a}, \tag{11}$$

where I_n is the $n \times n$ identity matrix. Indeed, without loss of generality, we can let $\boldsymbol{\Sigma}$ be of the form $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^0 + \rho\Delta\boldsymbol{\Sigma}$. We then obtain

$$\max_{\boldsymbol{\Sigma}\,:\,\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^0\|_F\leq\rho} \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} = \mathbf{a}^T\boldsymbol{\Sigma}^0\mathbf{a} + \rho \max_{\Delta\boldsymbol{\Sigma}\,:\,\|\Delta\boldsymbol{\Sigma}\|_F\leq 1} \mathbf{a}^T\Delta\boldsymbol{\Sigma}\mathbf{a} = \mathbf{a}^T\boldsymbol{\Sigma}^0\mathbf{a} + \rho\mathbf{a}^T\mathbf{a}, \tag{12}$$

using the Cauchy-Schwarz inequality for

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

. This holds of compatibility of the Frobenius matrix norm and the euclidean vector norm and because $\|\Delta\boldsymbol{\Sigma}\|_F \leq 1$. For $\Delta\boldsymbol{\Sigma}$ the unity matrix, this upper bound is attained.

Using (7) and (11), we obtain the robust version of the core result (**??**):

$$\inf_{\mathbf{y}\sim(\bar{\mathbf{y}},\boldsymbol{\Sigma}_{\mathbf{y}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \alpha \forall(\bar{\mathbf{y}},\boldsymbol{\Sigma}_{\mathbf{y}}) \in \mathcal{Y} \Leftrightarrow b - \mathbf{a}^T\bar{\mathbf{y}} \geq (\kappa(\alpha)+\nu)\sqrt{\mathbf{a}^T(\boldsymbol{\Sigma}_{\mathbf{y}} + \rho I_n)\mathbf{a}} \quad \text{with} \quad \kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}. \tag{13}$$

Applying this result to (4) finally gives us that the optimal robust minimax probability classifier from problem (4) with sets $\mathcal{X}$, $\mathcal{Y}$ given by (5) can be obtained by solving problem (3), with $\boldsymbol{\Sigma}_{\mathbf{x}} = {\boldsymbol{\Sigma}_{\mathbf{x}}}^0 + \rho I_n$, $\boldsymbol{\Sigma}_{\mathbf{y}} = {\boldsymbol{\Sigma}_{\mathbf{y}}}^0 + \rho I_n$. If κ_*^{-1} is the optimal value of that problem, the corresponding worst-case misclassification probability is

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

So, with only uncertainty in the mean, the robust hyperplane is the *same* as the non-robust one; the only change is in the increase in the worst-case misclassification probability. The uncertainty in the covariance matrix adds a term ρI_n to the covariance matrices, which can be interpreted as regularization term. This will change the hyperplane and also increases the worst-case misclassification probability. Also notice that, if there is too much uncertainty in the mean (i.e. $\kappa_* < \nu$), the robust version is not feasible: there is no way to find a hyperplane which separates the two classes in the robust minimax probabilistic sense and the worst-case misclassification probability is $1 - \alpha_*^{\text{rob}} = 1$.

Just as for the non-robust case (see [a]), this robust approach can be generalized to allow nonlinear decision boundaries via the use of Mercer kernels. For results, we refer the reader to [a].

3 Single class case for robust novelty detection

In this Section we extend our minimax approach to classification to the single class case, and this directly in a robust setting as well.

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 $\mathbf{Pr}\{\mathbf{x} \in \mathcal{Q}\} = \alpha$ [b]. 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?

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 $\mathbf{\Sigma_x}$. We assume again that the real mean and covariance matrix are unknown but bounded in a set $\mathcal{X}$ as specified in (5):

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

We want the region $\mathcal{Q}$ to be tight, so we maximize its Mahalanobis distance to the origin. We do this in a robust setting: we maximize the Mahalanobis distance in the worst-case estimate/choice of the covariance matrix, i.e., the one that gives us the smallest Mahalanobis distance:

$$\max_{\mathbf{a} \neq 0, b} \min_{(\bar{\mathbf{x}}, \mathbf{\Sigma_x}) \in \mathcal{X}} \frac{b}{\sqrt{\mathbf{a}^T \mathbf{\Sigma_x} \mathbf{a}}} \quad \text{s.t.} \quad \inf_{\mathbf{x} \sim (\bar{\mathbf{x}}, \mathbf{\Sigma_x})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x} \geq b\} \geq \alpha \quad \forall (\bar{\mathbf{x}}, \mathbf{\Sigma_x}) \in \mathcal{X}. \tag{14}$$

For simplicity, we will assume that $\mathbf{\Sigma_x}$ is positive definite; our result is easily extended to the general positive semidefinite case. (nodig????? weg als nergens pos def explicit assumed wordt) Note that $\mathcal{Q}(\mathbf{a}, b)$ does not contain $\mathbf{0}$ if and only if $b > 0$. Also, the optimization problem (14) is positively homogeneous in $(\mathbf{a}, b)$. Thus, without loss of generality, we can set $b = 1$ in problem (14). Furthermore, we can use the robust core result (13) for the constraint in (14), with $\mathbf{a} \neq 0$, and obtain

$$\min_{\mathbf{a}} \max_{(\bar{\mathbf{x}}, \mathbf{\Sigma_x}) \in \mathcal{X}} \mathbf{a}^T \mathbf{\Sigma_x} \mathbf{a} \quad \text{s.t.} \quad \mathbf{a}^T\bar{\mathbf{x}} - 1 \geq (\kappa(\alpha) + \nu)\sqrt{\mathbf{a}^T(\mathbf{\Sigma_x} + \rho I_n)\mathbf{a}}, \tag{15}$$

where $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$ and the $\mathbf{a} \neq 0$ can be omitted since the constrained wouldn't be satisfied in this case. Using (11), this can be written as

$$\min_{\mathbf{a}} \mathbf{a}^T(\mathbf{\Sigma_x} + \rho I_n)\mathbf{a} \quad \text{s.t.} \quad \mathbf{a}^T\bar{\mathbf{x}} - 1 \geq (\kappa(\alpha) + \nu)\sqrt{\mathbf{a}^T(\mathbf{\Sigma_x} + \rho I_n)\mathbf{a}}, \tag{16}$$

Thus an optimal half-space $\mathcal{Q}(\mathbf{a}, b)$ can be determined again by solving a (convex) second order cone programming problem. The worst-case probability of occurrence outside region $\mathcal{Q}$ is then given by $1-\alpha$. Notice that the particular choice of $\alpha \in (0,1]$ must be feasible, i.e.,

$$\exists\, \mathbf{a} \ : \ \mathbf{a}^T\bar{\mathbf{x}} - 1 \geq (\kappa(\alpha) + \nu)\sqrt{\mathbf{a}^T(\mathbf{\Sigma_x} + \rho I_n)\mathbf{a}}.$$

In this case, $\mathbf{\Sigma_x} + \rho I_n$ is positive definite and the halfspace is unique. Even more, it can explicitly be determined. Indeed, (**??**) can also be written as:

$$\begin{aligned}\min_{\mathbf{a}}\ \mathbf{a}^T(\mathbf{\Sigma_x}+\rho I_n)\mathbf{a} \qquad \text{s.t.} \qquad & (\mathbf{a}^T\bar{\mathbf{x}}-1)^2 \geq (\kappa(\alpha)+\nu)^2(\mathbf{a}^T(\mathbf{\Sigma_x}+\rho I_n)\mathbf{a}) \\ & \mathbf{a}^T\bar{\mathbf{x}}-1 \geq 0. \qquad (17)\end{aligned}$$

To solve this, we form the Lagrangian

$$\begin{aligned}\mathcal{L}(\mathbf{a},\lambda,\sigma) &= \mathbf{a}^T(\mathbf{\Sigma_x}+\rho I_n)\mathbf{a} + \lambda\left[(\kappa(\alpha)+\nu)^2(\mathbf{a}^T(\mathbf{\Sigma_x}+\rho I_n)\mathbf{a}) - (\mathbf{a}^T\bar{\mathbf{x}}-1)^2\right] - \sigma(\mathbf{a}^T\bar{\mathbf{x}}-1) \\ &= \mathbf{a}^T\left[(1+\lambda(\kappa(\alpha)+\nu)^2)(\mathbf{\Sigma_x}+\rho I_n) - \lambda\bar{\mathbf{x}}\bar{\mathbf{x}}^T\right]\mathbf{a} - (\sigma-2\lambda)\bar{\mathbf{x}}^T\mathbf{a} - \lambda + \sigma,\end{aligned}$$

which is to be maximized with respect to $\lambda \geq 0, \sigma \geq 0$ and minimized with respect to $\mathbf{a}$. At the optimum, we have $\frac{\partial}{\partial\mathbf{a}}\mathcal{L}(\mathbf{a},\lambda,\sigma) = 0$, which boils down to:

$$\begin{aligned}\left[(1+\lambda(\kappa(\alpha)+\nu)^2)(\mathbf{\Sigma_x}+\rho I_n) - \lambda\bar{\mathbf{x}}\bar{\mathbf{x}}^T\right]\mathbf{a} &= (\sigma-2\lambda)\bar{\mathbf{x}} \\ \Rightarrow (1+\lambda(\kappa(\alpha)+\nu)^2)(\mathbf{\Sigma_x}+\rho I_n)\mathbf{a} &= (\sigma-2\lambda+\lambda\bar{\mathbf{x}}^T\mathbf{a})\bar{\mathbf{x}} \\ \Rightarrow \mathbf{a} &= \frac{\sigma-2\lambda+\lambda\bar{\mathbf{x}}^T\mathbf{a}}{1+\lambda(\kappa(\alpha)+\nu)^2}(\mathbf{\Sigma_x}+\rho I_n)^{-1}\bar{\mathbf{x}} \quad (18)\end{aligned}$$

This implies that the optimal $\mathbf{a}$ is proportional to $(\mathbf{\Sigma_x}+\rho I_n)^{-1}\bar{\mathbf{x}}$. Hence, we can write $\mathbf{a}$ as $\tau(\mathbf{\Sigma_x}+\rho I_n)^{-1}\bar{\mathbf{x}}$ where $\tau \geq 0$ because otherwise $\mathbf{a}^T\mathbf{x} \leq 0$ and the second constraint in (17) cannot be satisfied. When substituting $\tau(\mathbf{\Sigma_x}+\rho I_n)^{-1}\bar{\mathbf{x}}$ for $\mathbf{a}$ in (**??**), we get

$$\min_{\tau\geq 0}\ \tau^2\zeta^2 \quad \text{s.t.} \quad \tau\zeta^2 - 1 \geq (\kappa(\alpha)+\nu)\tau\zeta, \qquad (19)$$

where $\zeta^2 = \bar{\mathbf{x}}^T(\mathbf{\Sigma_x}+\rho I_n)^{-1}\bar{\mathbf{x}} > 0$ because $\bar{\mathbf{x}} \neq \mathbf{0}$ (implied by the second constraint in (17)). The constraint can be written as $\tau(\zeta^2 - (\kappa(\alpha)+\nu)\zeta) \geq 1$. Because $\tau \geq 0$, this can only be satisfied if $\zeta^2 - (\kappa(\alpha)+\nu)\zeta \geq 0$, which is nothing other than the feasibility condition for α:

$$\begin{aligned}\zeta^2 - (\kappa(\alpha)+\nu)\zeta \geq 0 \quad &\Leftrightarrow \quad \kappa(\alpha) \leq \zeta - \nu \\ &\Leftrightarrow \quad \alpha \leq \frac{(\zeta-\nu)^2}{1+(\zeta-\nu)^2}.\end{aligned}$$

If this is fulfilled, the optimization (19) is feasible and boils down to

$$\min_{\tau\geq 0}\ \tau \quad \text{s.t.} \quad \tau \geq \frac{1}{\zeta^2 - (\kappa(\alpha)+\nu)\zeta}.$$

One can easily see that the optimal τ is given by $\tau_* = 1/(\zeta^2 - (\kappa(\alpha)+\nu)\zeta)$ which leads to

$$\mathbf{a}_* = \frac{\mathbf{\Sigma_x}^{-1}\bar{\mathbf{x}}}{\zeta^2 - \kappa(\alpha)\zeta} \qquad \text{with} \quad \zeta = \sqrt{\bar{\mathbf{x}}^T\mathbf{\Sigma_x}^{-1}\bar{\mathbf{x}}}, \qquad (20)$$

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 $\bar{\mathbf{x}} = \mathbf{0}$ (in this case, no $\alpha \in (0,1]$ will be feasible), problem (14) does not have a solution.

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 $1-\alpha$.

In a similar way as for the binary classification problem (see [a]), one can obtain a nonlinear region $\mathcal{Q}$ in $\mathbb{R}^n$ for the single class case, by mapping the data $\mathbf{x} \mapsto \varphi(\mathbf{x}) \sim (\overline{\varphi(\mathbf{x})}, \mathbf{\Sigma}_{\varphi(\mathbf{x})})$ and solving (14) in the higher-dimensional feature space $\mathbb{R}^f$:

$$\max_{\mathbf{a}\neq 0,b}\ \frac{b}{\sqrt{\mathbf{a}^T\mathbf{\Sigma}_{\varphi(\mathbf{x})}\mathbf{a}}} \qquad \text{s.t.} \qquad \inf_{\varphi(\mathbf{x})\sim(\overline{\varphi(\mathbf{x})},\mathbf{\Sigma}_{\varphi(\mathbf{x})})} \mathbf{Pr}\{\mathbf{a}^T\varphi(\mathbf{x}) \geq b\} \geq \alpha. \qquad (21)$$

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. We finally obtain that, if the choice of α is feasible, that is

$$\exists\, \gamma \;:\; \gamma^T \tilde{\mathbf{k}}_{\mathbf{x}} - 1 \geq (\kappa(\alpha) + \nu)\sqrt{\gamma^T (\tilde{\mathbf{K}}_{\mathbf{x}}^T \tilde{\mathbf{K}}_{\mathbf{x}} + \rho \mathbf{K})\gamma},$$

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

$$\min_{\gamma}\; \gamma^T (\tilde{\mathbf{K}}_{\mathbf{x}}^T \tilde{\mathbf{K}}_{\mathbf{x}} + \rho \mathbf{K})\gamma \quad \text{s.t.} \quad \gamma^T \tilde{\mathbf{k}}_{\mathbf{x}} - 1 \geq (\kappa(\alpha) + \nu)\sqrt{\gamma^T (\tilde{\mathbf{K}}_{\mathbf{x}}^T \tilde{\mathbf{K}}_{\mathbf{x}} + \rho \mathbf{K})\gamma},$$

where $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$ and $b = 1$. where $\gamma = [\alpha_1\ \alpha_2\ \cdots\ \alpha_{N_x}\ \beta_1\ \beta_2\ \cdots\ \beta_{N_y}]^T$, $\tilde{\mathbf{k}}_{\mathbf{x}} \in \mathbb{R}^{N_x+N_y}$ with $[\tilde{\mathbf{k}}_{\mathbf{x}}]_i = \frac{1}{N_x}\sum_{j=1}^{N_x} K(\mathbf{x}_j, \mathbf{z}_i)$, $\tilde{\mathbf{k}}_{\mathbf{y}} \in \mathbb{R}^{N_x+N_y}$ with $[\tilde{\mathbf{k}}_{\mathbf{y}}]_i = \frac{1}{N_y}\sum_{j=1}^{N_y} K(\mathbf{y}_j, \mathbf{z}_i)$, $\mathbf{z}_i = \mathbf{x}_i$ for $i = 1, 2, \ldots, N_x$ and $\mathbf{z}_i = \mathbf{y}_{i-N_x}$ for $i = N_x + 1, N_x + 2, \ldots, N_x + N_y$. $\tilde{\mathbf{K}}$ is defined as:

$$\tilde{\mathbf{K}} = \begin{pmatrix} \mathbf{K}_{\mathbf{x}} - \mathbf{1}_{N_x}\tilde{\mathbf{k}}_{\mathbf{x}}^T \\ \mathbf{K}_{\mathbf{y}} - \mathbf{1}_{N_y}\tilde{\mathbf{k}}_{\mathbf{y}}^T \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{K}}_{\mathbf{x}} \\ \tilde{\mathbf{K}}_{\mathbf{y}} \end{pmatrix} \tag{22}$$

where $\mathbf{1}_m$ is a column vector with ones of dimension m. $\mathbf{K}_{\mathbf{x}}$ and $\mathbf{K}_{\mathbf{y}}$ contain respectively the first N_x rows and the last N_y rows of the Gram matrix $\mathbf{K}$ (defined as $\mathbf{K}_{ij} = \varphi(\mathbf{z}_i)^T \varphi(\mathbf{z}_j) = K(\mathbf{z}_i, \mathbf{z}_j)$).

The worst-case probability of occurrence outside region $\mathcal{Q}$ is given by $1 - \alpha$. If $\tilde{\mathbf{K}}_{\mathbf{x}}^T \tilde{\mathbf{K}}_{\mathbf{x}} + \rho \mathbf{K}$ is positive definite, the optimal half-space is unique and determined by

$$\gamma_* = \frac{(\tilde{\mathbf{K}}_{\mathbf{x}}^T \tilde{\mathbf{K}}_{\mathbf{x}} + \rho \mathbf{K})^{-1}\tilde{\mathbf{k}}_{\mathbf{x}}}{\zeta^2 - (\kappa(\alpha) + \nu)\zeta} \qquad \text{with} \quad \zeta = \sqrt{\tilde{\mathbf{k}}_{\mathbf{x}}^T (\tilde{\mathbf{K}}_{\mathbf{x}}^T \tilde{\mathbf{K}}_{\mathbf{x}})^{-1}\tilde{\mathbf{k}}_{\mathbf{x}}},$$

if the choice of α is such that $\kappa(\alpha) \leq \zeta - \nu$ or $\alpha \leq \frac{(\zeta-\nu)^2}{1+(\zeta-\nu)^2}$. When we numerically regularize $\tilde{\mathbf{K}}_{\mathbf{x}}^T \tilde{\mathbf{K}}_{\mathbf{x}} + \rho \mathbf{K}$ with an extra term ϵI_n, this unique solution can always be determined.

If the choice of α is not feasible or if $\tilde{\mathbf{k}}_{\mathbf{x}} = \mathbf{0}$ (in this case, no $\alpha \in (0, 1]$ will be feasible), problem (21) 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 $1 - \alpha$.

4 Experiments

"exp and discussion"

Their values can be determined based on resampling methods, or, for a specific application (e.g., binary classification, single class), they can be considered as hyperparameters of the algorithm and tuned using e.g., cross-validation.

In this section we report empirical results that test our algorithmic approach, the robust single class MPM, and compare it to the single class SVM (see [c]) and to a two-class SVM approach where an artificial second class is obtained by uniformly generating data points in $T = \{\mathbf{z} \in \mathbb{R}^n \,|\, \min\{(\mathbf{x}_1)_i, (\mathbf{x}_2)_i, \ldots, (\mathbf{x}_n)_i\} \leq \mathbf{z}_i \leq \max\{(\mathbf{x}_1)_i, (\mathbf{x}_2)_i, \ldots, (\mathbf{x}_n)_i\}\}$.

For training, we fed the different algortihms with 80% of the data points of one class of different standard benchmark data sets, using a Gaussian kernel of width σ ($e^{-\|x-y\|^2/\sigma}$). The other 20% of those data points and the data points of the other class are then used for testing the performance. The Wisconsin breast cancer dataset contained 16 missing examples which were not used. The breast cancer and sonar data were obtained from the UCI repository while the heart data were obtained from STATLOG and normalized. Data for the twonorm problem data were generated as specified in [3]. The kernel parameter (σ) for the Gaussian kernel was tuned using cross-validation over 20 random partitions, as was the parameter ρ for $\mathcal{X}$. For simplicity, we put $\nu = 0$ for the single class MPM (this parameter doesn't really influence the performance anyway, it only decreases the actual value of α) The averages over 30 random partitions into 80% training set and 20% test set are reported in Table 1. The desired trade-off between a good false positive and false negative performance can be obtained be tuning α, ν (fraction of support vectors and outliers for single class SVM) or C (SVM soft margin weight parameter) appropriately.

Table 1: Performance of different algorithms for single class problems; the better performance per line is indicated in italic; FP = false positives (out-of-class data not detected as such); FN = false negatives (in-class-data not detected as such).

Dataset	*Single Class MPM*			*Single Class SVM*			*Two-Class SVM approach*		
	α	FP	FN	ν	FP	FN	C	FP	FN
Sonar	0.2	*24.7 %*	*64.0 %*	0.6	26.9 %	65.4 %	0.1	23.8 %	68.6 %
class +1	0.8	*44.6 %*	*39.6 %*	0.2	47.3 %	42.1 %	0.2	48.3 %	42.3 %
	0.95	69.3 %	17.3 %	0.0005	75.4 %	16.2 %	1	*75.2 %*	*16.0 %*
Sonar	0.6	5.4 %	51.7 %	0.4	8.5 %	53.7 %	0.1	9.7 %	70.0 %
class -1	0.9	10.0 %	37.4 %	0.001	15.7 %	41.3 %	0.2	34.6 %	40.6 %
	0.95	19.1 %	29.7 %	0.0006	36.1 %	28.4 %	0.35	47.7 %	26.0 %
	0.99	56.1 %	5.7 %	0.0003	82.6 %	6.3 %	1	67.9 %	6.1 %
Breast	0.6	0.0 %	8.8 %	0.14	0.0 %	14.6 %	0.005	0.4 %	8.0 %
Cancer	0.8	1.8 %	5.9 %	0.001	2.4 %	6.1 %	0.1	0.9 %	4.3 %
class +1	0.2	10.5 %	2.7 %	0.0003	11.5 %	3.1 %	10	12.3 %	3.1 %
Breast	0.01	2.4 %	26.5 %	0.4	2.5 %	41.4 %	0.8	0.9 %	47.9 %
Cancer	0.03	2.9 %	13.5 %	0.2	2.8 %	25.0 %	1	11.0 %	45 %
class -1	0.05	3.0 %	8.3 %	0.1	3.1 %	11.3 %	2	89.2 %	38.2 %
	0.14	5.9 %	1.9 %	0.0005	9.2 %	3.4 %	100	98.0 %	23.5 %
Twonorm	0.01	6.3 %	43.2 %	0.4	6.2 %	42.8 %	0.13	6.8 %	37.3 %
class +1	0.2	13.9 %	22.5 %	0.2	12.7 %	22.8 %	0.17	12.0 %	24.2 %
	0.4	22.5 %	11.9 %	0.0008	23.3 %	9.6 %	5	25.9 %	10.5 %
	0.6	36.9 %	4.5 %	0.0003	33.4 %	4.5 %			
Twonorm	0.1	5.6 %	43.7 %	0.4	6.0 %	44.1 %	0.35	6.1 %	49.8 %
class -1	0.4	11.3 %	23.1 %	0.15	11.8 %	24.6 %	0.5	24.5 %	23.7 %
	0.6	16.9 %	12.1 %	0.0005	35.9 %	12.0 %	10	30.1 %	10.0 %
	0.8	30.1 %	6.9 %	0.0003	39.3 %	6.9 %			
Heart	0.46	13.4 %	46.2 %	0.4	13.5 %	47.8 %	0.05	11.9 %	46.4 %
class +1	0.52	24.0 %	30.9 %	0.05	24.8 %	36.7 %	0.07	22.1 %	30.3 %
	0.54	33.5 %	22.6 %	0.0008	38.8 %	27.0 %	0.1	35.8 %	22.9 %
Heart	0.0001	15.9 %	41.3 %	0.4	20.8 %	50.7 %	0.08	13.9 %	43.8 %
class -1	0.0006	21.2 %	37.2 %	0.002	26.3 %	43.8 %	0.09	21.0 %	37.5 %
	0.003	36.3 %	27.2 %	0.0007	43.7 %	29.2 %	0.11	39.2 %	31.8 %
	0.01	56.9 %	15.9 %	0.0005	58.4 %	18.09 %	0.2	68.6 %	16.7 %

First of all, notice that the average performance of the single class MPM is competitive with or even better than the best of the other approaches for the above datasets. This certainly justifies the minimax approach to novelty detection from an empirical point of view. However, the single class SVM and the 2-class approach also show to be valuable, although the latter sometimes fails to provide us with an extensive range of trade-off between false positives and false negatives.

Also, notice that $1-\alpha$ (worst-case probability of false negatives) is indeed an upper bound on the average percentage of false negatives, except for the class -1 of the sonar data set. The latter indicates that the simplifying assumption $\nu = 0$ is false in this case.

It's also interesting to see how the performance of the single class MPM is affected when, for a given α, ρ is being put to zero, meaning no uncertainty on the covariance estimate is assumed. Without giving any numerical results, we can tell here that putting $\rho = 0$ indeed deteriorated the performance, usually giving a rather low false positive rate but a very bad false negative rate.

In these experiments, we put $\nu = 0$ for simplicity and considered ρ rather as a general hyperparameter of our model (a knob, being optimally tuned using cross-validation) then as an inherent statistical parameter. Although this approach is perfectly justifiable, it makes the bound $1-\alpha$ statistically less significant than when estimating ν and ρ from data to the best extent possible (e.g. using resampling). A hidden detail however is that ν and ρ will depend on σ, to be determined using cross-validation anyway.

5 Conclusions

After making the MPM formulation robust against estimation errors in the means and covariance matrices, the robust minimax approach is naturally extended towards the unsupervised learning problem of quantile estimation. Positive empirical results on standard benchmark datasets are obtained and support the use of the single class MPM.

Acknowledgements

We would like to acknowledge support from ONR MURI N00014-00-1-0637. Thanks to Alex Smola for constructive conversations and suggestions.

References

[1] Ben-David, S. and Lindenbaum, M. Learning distributions by their density levels: A paradigm for learning without a teacher. *Journal of Computer and System Sciences*, 55:171-182, 1997.

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

[3] Lanckriet, G., El Ghaoui, L., Bhattacharyya, C. and Jordan, M. (2001) Minimax Probability Machine. Submitted to *Journal of Machine Learning Research.*

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

[5] Tax, D. and Duin, R. (1999) Data domain description by support vectors. In *Proceedings ESANN*, pp. 251-256, Brussels, 1999.