Journal of Machine Learning Research 5 (2004) 1253–1286 Submitted 8/03; Revised 4/04; Published 10/04

The Minimum Error Minimax Probability Machine

Kaizhu Huang KZHUANG@CSE.CUHK.EDU.HK
Haiqin Yang HQYANG@CSE.CUHK.EDU.HK
Irwin King KING@CSE.CUHK.EDU.HK
Michael R. Lyu LYU@CSE.CUHK.EDU.HK
Laiwan Chan LWCHAN@CSE.CUHK.EDU.HK
Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, N. T., Hong Kong

Editor: Michael I. Jordan

Abstract

We construct a distribution-free Bayes optimal classifier called the Minimum Error Minimax Probability Machine (MEMPM) in a worst-case setting, i.e., under all possible choices of class-conditional densities with a given mean and covariance matrix. By assuming no specific distributions for the data, our model is thus distinguished from traditional Bayes optimal approaches, where an assumption on the data distribution is a must. This model is extended from the Minimax Probability Machine (MPM), a recently-proposed novel classifier, and is demonstrated to be the general case of MPM. Moreover, it includes another special case named the Biased Minimax Probability Machine, which is appropriate for handling biased classification. One appealing feature of MEMPM is that it contains an explicit performance indicator, i.e., a lower bound on the worst-case accuracy, which is shown to be tighter than that of MPM. We provide conditions under which the worst-case Bayes optimal classifier converges to the Bayes optimal classifier. We demonstrate how to apply a more general statistical framework to estimate model input parameters robustly. We also show how to extend our model to nonlinear classification by exploiting kernelization techniques. A series of experiments on both synthetic data sets and real world benchmark data sets validates our proposition and demonstrates the effectiveness of our model.

Keywords: classification, distribution-free, kernel, minimum error, sequential biased minimax probability machine, worst-case accuracies

1. Introduction

A novel model for two-category classification tasks called the Minimax Probability Machine (MPM) has been recently proposed (Lanckriet et al., 2002a). This model tries to minimize the probability of misclassification of future data points in a worst-case setting, i.e., under all possible choices of class-conditional densities with a given mean and covariance matrix. When compared with traditional generative models, MPM avoids making assumptions with respect to the data distribution; such assumptions are often invalid and lack generality. This model's performance is reported to be comparable to the Support Vector Machine (SVM) (Vapnik, 1999), a state-of-the-art classifier.

However, MPM forces the worst-case accuracies for two classes to be equal. This constraint seems inappropriate, since it is unnecessary that the worst-case accuracies are presumed equal.

Therefore, the classifier derived from this model does not result in minimizing the worst-case error rate of future data points and thus in a sense cannot represent the optimal classifier.

In this paper, by removing this constraint, we propose a generalized Minimax Probability Machine, called the Minimum Error Minimax Probability Machine (MEMPM). Instead of optimizing an equality-constrained worst-case error rate, this model minimizes the worst-case Bayes error rate of future data and thus achieves the optimum classifier in the worst-case scenario. Furthermore, this new model contains several appealing features.

First, as a generalized model, MEMPM includes and expands the Minimax Probability Machine. Interpretations from the viewpoints of the optimal thresholding problem and the geometry will be provided to show the advantages of MEMPM. Moreover, this generalized model includes another promising special case, named the Biased Minimax Probability Machine (BMPM) (Huang et al., 2004b), and extends its application to a type of important classification, i.e., biased classification.

Second, this model derives a distribution-free Bayes optimal classifier in the worst-case scenario. It thus distinguishes itself from the traditional Bayes optimal classifiers, which have to assume distributions for the data and thus lack generality in real cases. Furthermore, we will show that, under certain conditions, e.g., when a Gaussian distribution is assumed for the data, the worst-case Bayes optimal classifier becomes the true Bayes optimal hyperplane.

Third, similar to MPM, the MEMPM model also contains an explicit performance indicator, namely an explicit upper bound on the probability of misclassification of future data. Moreover, we will demonstrate theoretically and empirically that MEMPM attains a smaller upper bound of the probability of misclassification than MPM, which thus implies the superiority of MEMPM to MPM.

Fourth, although in general the optimization of MEMPM is shown to be a non-concave problem, empirically, it demonstrates reasonable concavity in the main "interest" region and thus can be solved practically. Furthermore, we will show that the final optimization problem involves solving a one-dimensional line search problem and thus results in a satisfactory solution.

This paper is organized as follows. In the next section, we present the main content of this paper, the MEMPM model, including its definition, interpretations, practical solving method, and sufficient conditions for convergence to the true Bayes decision hyperplane. Following that, we demonstrate a robust version of MEMPM. In Section 4, we seek to kernelize the MEMPM model to attack nonlinear classification problems. We then, in Section 5, present a series of experiments on synthetic data sets and real world benchmark data sets. In Section 6, we analyze the tightness of the worst-case accuracy bound. In Section 7, we show that empirically MEMPM is often concave in the main "interest" region. In Section 8, we present the limitations of MEMPM and envision possible future work. Finally, we conclude this paper in Section 9.

2. Minimum Error Minimax Probability Decision Hyperplane

In this section, we first present the model definition of MEMPM while reviewing the original MPM model. We then in Section 2.2 interpret MEMPM with respect to MPM. In Section 2.3, we specialize the MEMPM model for dealing with biased classification. In Section 2.4, we analyze the MEMPM optimization problem and propose a practical solving method. In Section 2.5, we address the sufficient conditions under which the worst-case Bayes optimal classifier derived from MEMPM becomes the true Bayes optimal classifier. In Section 2.6, we provide a geometrical interpretation for BMPM and MEMPM.

2.1 Problem Definition

The notation in this paper will largely follow that of Lanckriet et al. (2002b). Let $\mathbf{x}$ and $\mathbf{y}$ denote two random vectors representing two classes of data with means and covariance matrices as $\{\overline{\mathbf{x}}, \Sigma_{\mathbf{x}}\}$ and $\{\overline{\mathbf{y}}, \Sigma_{\mathbf{y}}\}$, respectively, in a two-category classification task, where $\mathbf{x}$, $\mathbf{y}$, $\overline{\mathbf{x}}$, $\overline{\mathbf{y}} \in \mathbb{R}^n$, and $\Sigma_{\mathbf{x}}$, $\Sigma_{\mathbf{y}} \in \mathbb{R}^{n\times n}$.

Assuming $\{\overline{\mathbf{x}}, \Sigma_{\mathbf{x}}\}$, $\{\overline{\mathbf{y}}, \Sigma_{\mathbf{y}}\}$ for two classes of data are reliable, MPM attempts to determine the hyperplane $\mathbf{a}^T\mathbf{z} = b$ ($\mathbf{a} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$, $\mathbf{z} \in \mathbb{R}^n$, $b \in \mathbb{R}$, and superscript T denotes the transpose) which can separate two classes of data with the maximal probability. The formulation for the MPM model is written as follows:

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

where α represents the lower bound of the accuracy for future data, namely, the worst-case accuracy. Future points $\mathbf{z}$ for which $\mathbf{a}^T\mathbf{z} \geq b$ are then classified as the class $\mathbf{x}$; otherwise they are judged as the class $\mathbf{y}$. This derived decision hyperplane is claimed to minimize the worst-case (maximal) probability of misclassification, or the error rate, of future data. Furthermore, this problem can be transformed to a convex optimization problem, or more specifically, a Second Order Cone Programming problem (Lobo et al., 1998; Nesterov and Nemirovsky, 1994).

As observed from the above formulation, this model assumes that the worst-case accuracies for two classes are the same. However, this assumption seems inappropriate, since it is unnecessary to require that the worst-case accuracies for two classes are exactly the same. Thus, the decision hyperplane given by this model does not necessarily minimize the worst-case error rate of future data and is not optimal in this sense. Motivated from the finding, we eliminate this constraint and propose a generalized model, the Minimum Error Minimax Probability Machine, as follows:

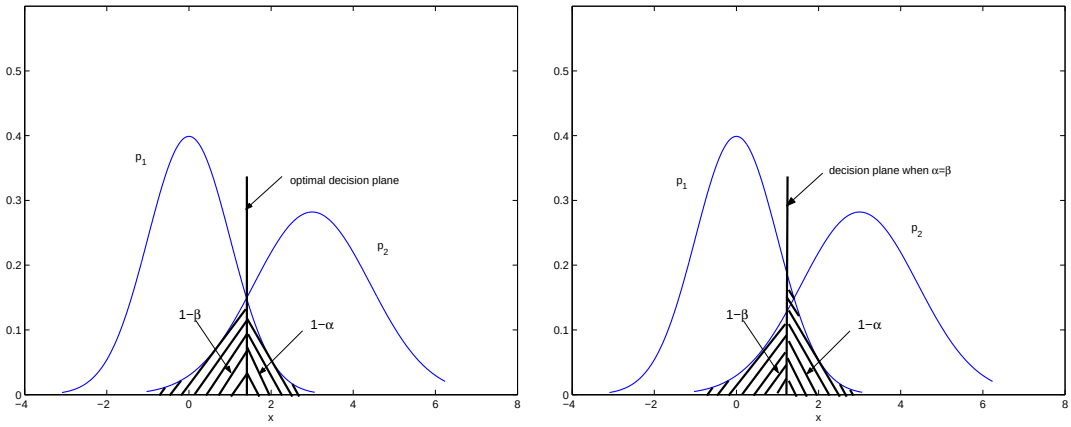
$$\begin{aligned} \max_{\alpha, \beta, \mathbf{a}\neq\mathbf{0}, b} \quad & \theta\alpha + (1-\theta)\beta \quad \text{s.t.} && (1) \\ & \inf_{\mathbf{x}\sim(\overline{\mathbf{x}}, \Sigma_{\mathbf{x}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x} \geq b\} \geq \alpha, && (2) \\ & \inf_{\mathbf{y}\sim(\overline{\mathbf{y}}, \Sigma_{\mathbf{y}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \beta. && (3) \end{aligned}$$

Similarly, α and β indicate the worst-case classification accuracies of future data points for the class $\mathbf{x}$ and $\mathbf{y}$, respectively, while $\theta \in [0,1]$ is the prior probability of the class $\mathbf{x}$ and $1-\theta$ is thus the prior probability of the class $\mathbf{y}$. Intuitively, maximizing $\theta\alpha + (1-\theta)\beta$ can naturally be considered as maximizing the expected worst-case accuracy for future data. In other words, this optimization leads to minimizing the expected upper bound of the error rate. More precisely, if we change $\max\{\theta\alpha + (1-\theta)\beta\}$ to $\min\{\theta(1-\alpha) + (1-\theta)(1-\beta)\}$ and consider $1-\alpha$ as the upper bound probability that an $\mathbf{x}$ data point is classified as the class $\mathbf{y}$ ($1-\beta$ is similarly considered), the MEMPM model exactly minimizes the maximum Bayes error and thus derives the Bayes optimal hyperplane in the worst-case scenario.

2.2 Interpretation

We interpret MEMPM with respect to MPM in this section. First, it is evident that if we presume $\alpha = \beta$, the optimization of MEMPM degrades to the MPM optimization. Therefore, MPM is a special case of MEMPM.

An analogy to illustrate the difference between MEMPM and MPM can be seen in the optimal thresholding problem. Figure 1 illustrates this analogy. To separate two classes of one-dimensional data with density functions as p_1 and p_2, respectively, the optimal thresholding is given by the decision plane in Figure 1(a) (assuming the prior probabilities for two classes of data are equal). This optimal thesholding corresponds to the point minimizing the error rate $(1-\alpha)+(1-\beta)$ or maximizing the accuracy $\alpha+\beta$, which is exactly the intersection point of two density functions ($1-\alpha$ represents the area of 135^o-line filled region and $1-\beta$ represents the area of 45^o-line filled region). On the other hand, the thresholding point to force $\alpha = \beta$ is not necessarily the optimal point to separate these two classes.



(a) (b)

Figure 1: An analogy to illustrate the difference between MEMPM and MPM with equal prior probabilities for two classes. The optimal decision plane corresponds to the intersection point, where the error $(1-\alpha)+(1-\beta)$ is minimized (or the accuracy $\alpha+\beta$ is maximized) as implied by MEMPM, rather than the one, where α is equal to β as implied by MPM.

It should be clarified that the MEMPM model assumes no distributions. This distinguishes the MEMPM model from the traditional Bayes optimal methods, which have to make specific assumptions on the data distribution. On the other hand, although MEMPM minimizes the upper bound of the Bayes error rate of future data points, as shown later in Section 2.5, it will represent the true Bayes optimal hyperplane under certain conditions, in particular, when Gaussianity is assumed for the data.

2.3 Special Case for Biased Classification

The above discussion only covers unbiased classification, which does not favor one class over the other class intentionally. However, another important type of pattern recognition tasks, namely biased classification, arises very often in practice. In this scenario, one class is usually more important than the other class. Thus a bias should be imposed towards the important class. Such typical example can be seen in the diagnosis of epidemical disease. Classifying a patient who is infected with a disease into the opposite class results in serious consequences. Thus in this problem, the classification accuracy should be biased towards the class with disease. In other words, we would prefer to diagnose the person without the disease to be the infected case rather than the other way round.

In the following we demonstrate that MEMPM contains a special case we call the Biased Minimax Probability Machine for biased classification. We formulate this special case as

$$\begin{aligned} \max_{\alpha,\beta,\mathbf{a}\neq\mathbf{0},b} \quad & \alpha \quad \text{s.t.} \\ & \inf_{\mathbf{x}\sim(\bar{\mathbf{x}},\Sigma_{\mathbf{x}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x} \geq b\} \geq \alpha, \\ & \inf_{\mathbf{y}\sim(\bar{\mathbf{y}},\Sigma_{\mathbf{y}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \beta_0, \end{aligned}$$

where $\beta_0 \in [0,\ 1)$, a pre-specified constant, represents an acceptable accuracy level for the less important class $\mathbf{y}$.

The above optimization utilizes a typical setting in biased classification, i.e., the accuracy for the important class (associated with $\mathbf{x}$) should be as high as possible, if only the accuracy for the less important class (associated with $\mathbf{y}$) maintains at an acceptable level specified by the lower bound β_0 (which can be set by users).

By quantitatively plugging a specified bias β_0 into classification and also by containing an explicit accuracy bound α for the important class, BMPM provides a direct and elegant means for biased classification. Comparatively, to achieve a specified bias, traditional biased classifiers such as the Weighted Support Vector Machine (Osuna et al., 1997) and the Weighted k-Nearest Neighbor method (Maloof et al., 2003) usually adapt different costs for different classes. However, due to the difficulties in establishing quantitative connections between the costs and the accuracy,[1] for imposing a specified bias, these methods have to resort to trial and error procedure to attain suitable costs; these procedures are generally indirect and lack rigorous treatments.

2.4 Solving the MEMPM Optimization Problem

In this section, we will propose to solve the MEMPM optimization problem. As will be demonstrated shortly, the MEMPM optimization can be transformed to a one-dimensional line search problem. More specifically, the objective function of the line search problem is implicitly determined by dealing with a BMPM problem. Therefore, solving the line search problem corresponds to solving a Sequential Biased Minimax Probability Machine (SBMPM) problem. Before we proceed, we first introduce how to solve the BMPM optimization problem.

1. Although cross validation might be used to provide empirical connections, they are problem-dependent and are usually slow procedures as well.

2.4.1 SOLVING THE BMPM OPTIMIZATION PROBLEM

First, we borrow Lemma 1 from Lanckriet et al. (2002b).

Lemma 1 *Given* $\mathbf{a} \neq \mathbf{0}$ *and b, such that* $\mathbf{a}^T\mathbf{y} \leq b$ *and* $\beta \in [0,1)$*, the condition*

$$\inf_{\mathbf{y}\sim(\overline{\mathbf{y}},\Sigma_\mathbf{y})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \beta,$$

holds if and only if $b - \mathbf{a}^T\overline{\mathbf{y}} \geq \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}}$ *with* $\kappa(\beta) = \sqrt{\frac{\beta}{1-\beta}}$.

By using Lemma 1, we can transform the BMPM optimization problem as follows:

$$\begin{aligned} \max_{\alpha,\mathbf{a}\neq\mathbf{0},b} \quad & \alpha \quad \text{s.t.} \\ & -b + \mathbf{a}^T\overline{\mathbf{x}} \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_\mathbf{x}\mathbf{a}}\,, \qquad (4) \\ & b - \mathbf{a}^T\overline{\mathbf{y}} \geq \kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}}\,, \qquad (5) \end{aligned}$$

where $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$, $\kappa(\beta_0) = \sqrt{\frac{\beta_0}{1-\beta_0}}$. (5) is directly obtained from (3) by using Lemma 1. Similarly, by changing $\mathbf{a}^T\mathbf{x} \geq b$ to $\mathbf{a}^T(-\mathbf{x}) \leq -b$, (4) can be obtained from (2).

From (4) and (5), we get

$$\mathbf{a}^T\overline{\mathbf{y}} + \kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}} \leq b \leq \mathbf{a}^T\overline{\mathbf{x}} - \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_\mathbf{x}\mathbf{a}}\,.$$

If we eliminate *b* from this inequality, we obtain

$$\mathbf{a}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}) \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_\mathbf{x}\mathbf{a}} + \kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}}\,. \qquad (6)$$

We observe that the magnitude of $\mathbf{a}$ does not influence the solution of (6). Moreover, we can assume $\overline{\mathbf{x}} \neq \overline{\mathbf{y}}$; otherwise, the minimax machine does not have a physical meaning. In this case, (6) may even have no solution for every $\beta_0 \neq 0$, since the right hand side would always be positive provided that $\mathbf{a} \neq \mathbf{0}$. Thus in the extreme case, α and β have to be zero, implying that the worst-case classification accuracy is always zero.

Without loss of generality, we can set $\mathbf{a}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}) = 1$. Thus the problem can further be changed to:

$$\begin{aligned} \max_{\alpha,\mathbf{a}\neq\mathbf{0}} \quad & \alpha \quad \text{s.t.} \\ & 1 \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_\mathbf{x}\mathbf{a}} + \kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}}\,, \qquad (7) \\ & \mathbf{a}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}) = 1. \end{aligned}$$

Since $\Sigma_\mathbf{x}$ can be assumed to be positive definite (otherwise, we can always add a small positive amount to its diagonal elements and make it positive definite), from (7) we can obtain:

$$\kappa(\alpha) \leq \frac{1 - \kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_\mathbf{x}\mathbf{a}}}\,.$$

Because $\kappa(\alpha)$ increases monotonically with α, maximizing α is equivalent to maximizing $\kappa(\alpha)$, which further leads to

$$\max_{\mathbf{a}\neq\mathbf{0}} \quad \frac{1-\kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_\mathbf{x}\mathbf{a}}} \quad \text{s.t.} \quad \mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}})=1.$$

This kind of optimization is called the Fractional Programming (FP) problem (Schaible, 1995). To elaborate further, this optimization is equivalent to solving the following fractional problem:

$$\max_{\mathbf{a}\neq\mathbf{0}} \quad \frac{f(\mathbf{a})}{g(\mathbf{a})}\,, \tag{8}$$

subject to $\mathbf{a}\in A=\{\mathbf{a}|\mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}})=1\}$, where $f(\mathbf{a})=1-\kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}}$, $g(\mathbf{a})=\sqrt{\mathbf{a}^T\Sigma_\mathbf{x}\mathbf{a}}$.

Theorem 2 *The Fractional Programming problem (8) associated with the BMPM optimization is a pseudo-concave problem, whose every local optimum is the global optimum.*

Proof It is easy to see that the domain A is a convex set on $\mathbb{R}^n$, and that $f(\mathbf{a})$ and $g(\mathbf{a})$ are differentiable on A. Moreover, since $\Sigma_\mathbf{x}$ and $\Sigma_\mathbf{y}$ can be both considered as positive definite matrices, $f(\mathbf{a})$ is a concave function on A and $g(\mathbf{a})$ is a convex function on A. Then $\frac{f(\mathbf{a})}{g(\mathbf{a})}$ is a concave-convex FP problem. Hence it is a pseudoconcave problem (Schaible, 1995). Therefore, every local maximum is the global maximum (Schaible, 1995). ■

To handle this specific FP problem, many methods such as the parametric method (Schaible, 1995), the dual FP method (Schaible, 1977; Craven, 1988), and the concave FP method (Craven, 1978) can be used. A typical Conjugate Gradient method (Bertsekas, 1999) in solving this problem has a worst-case $O(n^3)$ time complexity. Adding the time cost to estimate $\overline{\mathbf{x}}$, $\overline{\mathbf{y}}$, $\Sigma_\mathbf{x}$, and $\Sigma_\mathbf{y}$, the total cost for this method is $O(n^3+Nn^2)$, where N is the number of data points. This complexity is in the same order as the linear Support Vector Machines (Schölkopf and Smola, 2002) and the linear MPM (Lanckriet et al., 2002b).

In this paper, the Rosen gradient projection method (Bertsekas, 1999) is used to find the solution of this pseudo-concave FP problem, which is proved to converge to a local maximum with a worst-case linear convergence rate. Moreover, the local maximum will exactly be the global maximum in this problem.

2.4.2 SEQUENTIAL BMPM OPTIMIZATION METHOD FOR MEMPM

We now turn to solving the MEMPM problem. Similar to Section 2.4.1, we can base Lemma 1 to transform the MEMPM optimization as follows:

$$\max_{\alpha,\beta,\mathbf{a}\neq\mathbf{0},b} \quad \theta\alpha+(1-\theta)\beta \quad \text{s.t.}$$

$$-b+\mathbf{a}^T\overline{\mathbf{x}}\geq\kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_\mathbf{x}\mathbf{a}}\,, \tag{9}$$

$$b-\mathbf{a}^T\overline{\mathbf{y}}\geq\kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}}\,. \tag{10}$$

Using an analysis similar to that in Section 2.4.1, we can further transform the above optimization to:

$$\max_{\alpha,\beta,\mathbf{a}\neq\mathbf{0}} \quad \theta\alpha+(1-\theta)\beta \quad \text{s.t.} \tag{11}$$

$$1 \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}+\kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}\,, \tag{12}$$

$$\mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}})=1. \tag{13}$$

In the following we provide a lemma to show that the MEMPM solution is attained on the boundary of the set formed by the constraints of (12) and (13).

Lemma 3 *The maximum value of* $\theta\alpha+(1-\theta)\beta$ *under the constraints of (12) and (13) is achieved when the right hand side of (12) is strictly equal to* 1.

Proof Assume the maximum is achieved when $1>\kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}+\kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}$. A new solution constructed by increasing α or $\kappa(\alpha)$ a small positive amount,[2] and maintaining β, $\mathbf{a}$ unchanged will satisfy the constraints and will be a better solution. ■

By applying Lemma 3, we can transform the optimization problem (11) under the constraints of (12) and (13) as follows:

$$\max_{\beta,\mathbf{a}\neq\mathbf{0}} \quad \frac{\theta\kappa^2(\alpha)}{\kappa^2(\alpha)+1}+(1-\theta)\beta \quad \text{s.t.} \tag{14}$$

$$\mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}})=1, \tag{15}$$

where $\kappa(\alpha)=\frac{1-\kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}}$.

In (14), if we fix β to a specific value within $[0,1)$, the optimization is equivalent to maximizing $\frac{\kappa^2(\alpha)}{\kappa^2(\alpha)+1}$ and further equivalent to maximizing $\kappa(\alpha)$, which is exactly the BMPM problem. We can then update β according to some rules and repeat the whole process until an optimal β is found. This is also the so-called line search problem (Bertsekas, 1999). More precisely, if we denote the value of optimization as a function $f(\beta)$, the above procedure corresponds to finding an optimal β to maximize $f(\beta)$. Instead of using an explicit function as in traditional line search problems, the value of the function here is implicitly given by a BMPM optimization procedure.

Many methods can be used to solve the line search problem. In this paper, we use the Quadratic Interpolation (QI) method (Bertsekas, 1999). As illustrated in Figure 2, QI finds the maximum point by updating a three-point pattern $(\beta_1,\ \beta_2,\ \beta_3)$ repeatedly. The new β denoted by β_{new} is given by the quadratic interpolation from the three-point pattern. Then a new three-point pattern is constructed by β_{new} and two of β_1,β_2,β_3. This method can be shown to converge superlinearly to a local optimum point (Bertsekas, 1999). Moreover, as shown in Section 7, although MEMPM generally cannot guarantee its concavity, empirically it is often concave. Thus the local optimum will often be the global optimum in practice.

2. Since $\kappa(\alpha)$ increases monotonically with α, increasing α a small positive amount corresponds to increasing $\kappa(\alpha)$ a small positive amount.

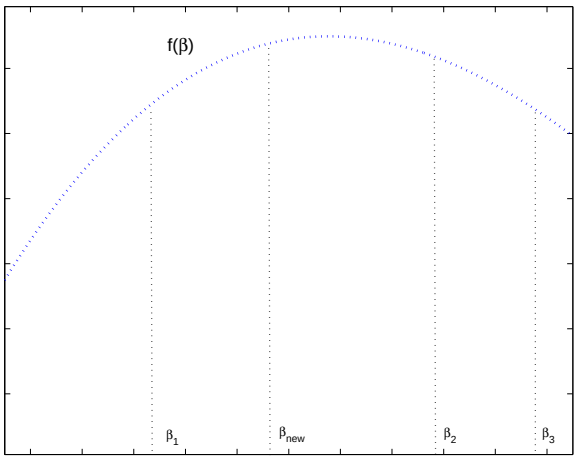


Figure 2: A three-point pattern and Quadratic Line search method. A β_{new} is obtained and a new three-point pattern is constructed by β_{new} and two of β_1, β_2 and β_3.

Until now, we do not mention how to calculate the intercept b. From Lemma 3, we can see that the inequalities (9) and (10) will become equalities at the maximum point $(\mathbf{a}_*, b_*)$. The optimal b will thus be obtained by

$$b_* = \mathbf{a}_*^T \overline{\mathbf{x}} - \kappa(\alpha_*)\sqrt{\mathbf{a}_*^T \Sigma_{\mathbf{x}} \mathbf{a}_*} = \mathbf{a}_*^T \overline{\mathbf{y}} + \kappa(\beta_*)\sqrt{\mathbf{a}_*^T \Sigma_{\mathbf{y}} \mathbf{a}_*}\,.$$

2.5 When Does the Worst-Case Bayes Optimal Hyperplane Become the True One?

As discussed, MEMPM derives the worst-case Bayes optimal hyperplane. Therefore, it is interesting to discover the conditions at which the worst-case optimal one changes to the true optimal one.

In the following we demonstrate two propositions. The first is that, when data are assumed to conform to some distributions, e.g., Gaussian distribution, the MEMPM framework leads to the Bayes optimal classifier; the second is that, when applied to high-dimensional classification tasks, the MEMPM model can be adapted to converge to the true Bayes optimal classifier under the Lyapunov condition.

To introduce the first proposition, we begin by assuming the data distribution as a Gaussian distribution.

Assuming $\mathbf{x} \sim \mathcal{N}(\overline{\mathbf{x}}, \Sigma_{\mathbf{x}})$ and $\mathbf{y} \sim \mathcal{N}(\overline{\mathbf{y}}, \Sigma_{\mathbf{y}})$, (2) becomes

$$\begin{aligned}
\inf_{\mathbf{x}\sim\mathcal{N}(\overline{\mathbf{x}},\Sigma_{\mathbf{x}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x} \geq b\} &= \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\overline{\mathbf{x}},\Sigma_{\mathbf{x}})}\{\mathbf{a}^T\mathbf{x} \geq b\} \\
&= \mathbf{Pr}\{\mathcal{N}(0,1) \geq \frac{b - \mathbf{a}^T\overline{\mathbf{x}}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}}\} \\
&= 1 - \Phi(\frac{b - \mathbf{a}^T\overline{\mathbf{x}}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}}) \\
&= \Phi(\frac{-b + \mathbf{a}^T\overline{\mathbf{x}}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}}) \geq \alpha, \qquad (16)
\end{aligned}$$

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

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

Due to the monotonic property of $\Phi(z)$, we can further write (16) as

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

Constraint (3) can be reformulated in a similar form. The optimization (1) is thus changed to:

$$\max_{\alpha,\beta,\mathbf{a}\neq\mathbf{0},b} \quad \theta\alpha + (1-\theta)\beta \quad \text{s.t.}$$

$$-b + \mathbf{a}^T \overline{\mathbf{x}} \geq \Phi^{-1}(\alpha) \sqrt{\mathbf{a}^T \Sigma_{\mathbf{x}} \mathbf{a}}\,, \tag{17}$$

$$b - \mathbf{a}^T \overline{\mathbf{y}} \geq \Phi^{-1}(\beta) \sqrt{\mathbf{a}^T \Sigma_{\mathbf{y}} \mathbf{a}}\,. \tag{18}$$

The above optimization is nearly the same as (1) subject to the constraints of (2) and (3) except that $\kappa(\alpha)$ is equal to $\Phi^{-1}(\alpha)$, instead of $\sqrt{\frac{\alpha}{1-\alpha}}$. Thus, it can similarly be solved based on the Sequential Biased Minimax Probability Machine method.

On the other hand, the Bayes optimal hyperplane corresponds to the one, $\mathbf{a}^T \mathbf{z} = b$ that minimizes the Bayes error:

$$\min_{\mathbf{a}\neq\mathbf{0},b} \quad \theta \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\overline{\mathbf{x}},\Sigma_{\mathbf{x}})}\{\mathbf{a}^T\mathbf{x} \leq b\} + (1-\theta)\mathbf{Pr}_{\mathbf{y}\sim\mathcal{N}(\overline{\mathbf{y}},\Sigma_{\mathbf{y}})}\{\mathbf{a}^T\mathbf{y} \geq b\}.$$

The above is exactly the upper bound of $\theta\alpha + (1-\theta)\beta$. From Lemma 3, we can know (17) and (18) will eventually become equalities. Traced back to (16), the equalities imply that α and β will achieve their upper bounds respectively. Therefore, when Gaussianity is assumed for the data, MEMPM derives the optimal Bayes hyperplane.

We propose Proposition 4 to extend the above analysis to general distribution assumptions.

Proposition 4 *If the distribution of the normalized random variable* $\frac{\mathbf{a}^T\mathbf{x} - \mathbf{a}^T\overline{\mathbf{x}}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}}$*, denoted as* $\mathcal{NS}$*, is independent of* $\mathbf{a}$*, minimizing the Bayes error bound in MEMPM exactly minimizes the true Bayes error, provided that* $\Phi(z)$ *is changed to* $\mathbf{Pr}\{\mathcal{NS}(0,1) \leq z\}$.

Before presenting Proposition 6, we first introduce the central limit theorem under the Lyapunov condition (Chow and Teicher, 1997).

Theorem 5 *Let* $\mathbf{x}_n$ *be a sequence of independent random variables defined on the same probability space. Assume that* $\mathbf{x}_n$ *has finite expected value* μ_n *and finite standard deviation* σ_n*. We define* $s_n^2 = \sum_{i=1}^n \sigma_i^2$*. Assume that the Lyapunov conditions are satisfied, namely, the third central moment* $r_n^3 = \sum_{i=1}^n \mathrm{E}(|\mathbf{x}_n - \mu_n|^3)$ *is finite for every n, and that* $\lim_{n\to\infty} \frac{r_n}{s_n} = 0$*. The sum* $S_n = \mathbf{x}_1 + \ldots + \mathbf{x}_n$ *converges towards a Gaussian distribution.*

One interesting finding directly elicited from the above central limit theorem is that, if the component variable $\mathbf{x}_i$ of a given n-dimensional random variable $\mathbf{x}$ satisfies the Lyapunov condition, the sum of weighted component variables $\mathbf{x}_i$, $1 \leq i \leq n$, namely, $\mathbf{a}^T\mathbf{x}$ tends towards a Gaussian distribution, as n grows.[3] This shows that, under the Lyapunov condition, when the dimension n grows,

3. Some techniques such as Independent Component Analysis (Deco and Obradovic, 1996) can be applied to decorrelate the dependence among random variables beforehand.

the hyperplane derived by MEMPM with the Gaussianity assumption tends towards the true Bayes optimal hyperplane. In this case, the MEMPM using $\Phi^{-1}(\alpha)$, the inverse function of the normal cumulative distribution, instead of $\sqrt{\frac{\alpha}{1-\alpha}}$, will converge to the true Bayes optimal decision hyperplane in the high-dimensional space. We summarize the analysis in Proposition 6.

Proposition 6 *If the component variable $\mathbf{x}_i$ of a given n-dimensional random variable $\mathbf{x}$ satisfies the Lyapunov condition, the MEMPM hyperplane derived by using $\Phi^{-1}(\alpha)$, the inverse function of the normal cumulative distribution, will converge to the true Bayes optimal one.*

The underlying justifications in the above two propositions are rooted in the fact that the generalized MPM is exclusively determined by the first and second moments. These two propositions emphasize the dominance of the first and second moments in representing data. More specifically, Proposition 4 hints that the distribution is only decided by up to the second moments. The Lyapunov condition in Proposition 6 also implies that the second order moment dominates the third order moment in the long run. It is also noteworthy that, with a fixed mean and covariance, the distribution of Maximum Entropy Estimation is a Gaussian distribution (Keysers et al., 2002). This would once again suggest the usage of $\Phi^{-1}(\alpha)$ in the high-dimensional space.

2.6 Geometrical Interpretation

In this section, we first provide a parametric solving method for BMPM. We then demonstrate that this parametric method enables a nice geometrical interpretation for both BMPM and MEMPM.

2.6.1 A PARAMETRIC METHOD FOR BMPM

We present a parametric method to solve BMPM in the following. When compared with Gradient methods, this approach is relatively slow, but it need not calculate the gradient in each step and hence may avoid accumulated errors.

According to the parametric method, the fractional function can be iteratively optimized in two steps (Schaible, 1995):

Step 1: Find $\mathbf{a}$ by maximizing $f(\mathbf{a})-\lambda g(\mathbf{a})$ in the domain A, where $\lambda\in\mathbb{R}$ is the newly introduced parameter.

Step 2: Update λ by $\frac{f(\mathbf{a})}{g(\mathbf{a})}$.

The iteration of the above two steps will guarantee to converge to a local maximum, which is also the global maximum in our problem. In the following, we adopt a method to solve the maximization problem in Step 1. Replacing $f(\mathbf{a})$ and $g(\mathbf{a})$, we expand the optimization problem to:

$$\max_{\mathbf{a}\neq\mathbf{0}} \quad 1-\kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}-\lambda\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}} \quad \text{s.t.} \quad \mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}})=1. \tag{19}$$

Maximizing (19) is equivalent to $\min_{\mathbf{a}}\kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}+\lambda\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}$ under the same constraint. By writing $\mathbf{a}=\mathbf{a}_0+\mathbf{F}\mathbf{u}$, where $\mathbf{a}_0=(\overline{\mathbf{x}}-\overline{\mathbf{y}})/\parallel\overline{\mathbf{x}}-\overline{\mathbf{y}}\parallel_2^2$ and $\mathbf{F}\in\mathbb{R}^{n\times(n-1)}$ is an orthogonal matrix whose columns span the subspace of vectors orthogonal to $\overline{\mathbf{x}}-\overline{\mathbf{y}}$, an equivalent form (a factor $\frac{1}{2}$ over each term has been dropped) to remove the constraint can be obtained:

$$\min_{\mathbf{u},\eta>0,\xi>0}\eta+\frac{\lambda^2}{\eta}\|{\Sigma_{\mathbf{x}}}^{1/2}(\mathbf{a}_0+\mathbf{F}\mathbf{u})\|_2^2+\xi+\frac{\kappa(\beta_0)^2}{\xi}\|{\Sigma_{\mathbf{y}}}^{1/2}(\mathbf{a}_0+\mathbf{F}\mathbf{u})\|_2^2\,,$$

where $\eta, \xi \in \mathbb{R}$. This optimization form is very similar to the one in the Minimax Probability Machine (Lanckriet et al., 2002a) and can also be solved by using an iterative least-squares approach.

2.6.2 A GEOMETRICAL INTERPRETATION FOR BMPM AND MEMPM

The parametric method enables a nice geometrical interpretation of BMPM and MEMPM in a fashion similar to that of MPM in Lanckriet et al. (2002b). Again, we assume $\overline{\mathbf{x}} \neq \overline{\mathbf{y}}$ for the meaningful classification and assume that $\Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{y}}$ are positive definite for the purpose of simplicity.

By using the 2-norm definition of a vector $\mathbf{z} : \|\mathbf{z}\|_2 = \max\{\mathbf{u}^T\mathbf{z} : \|\mathbf{u}\|_2 \leq 1\}$, we can express (19) as its dual form:

$$\tau_* := \min_{\mathbf{a}\neq\mathbf{0}} \max_{\mathbf{u},\mathbf{v}} \quad \lambda\mathbf{u}^T\Sigma_{\mathbf{x}}^{1/2}\mathbf{a} + \kappa(\beta_0)\mathbf{v}^T\Sigma_{\mathbf{y}}^{1/2}\mathbf{a} + \tau(1 - \mathbf{a}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}})) : \ \|\mathbf{u}\|_2 \leq 1, \|\mathbf{v}\|_2 \leq 1 \ .$$

We change the order of the min and max operators and consider the min:

$$\begin{aligned} &\min_{\mathbf{a}\neq\mathbf{0}} \quad \lambda\mathbf{u}^T\Sigma_{\mathbf{x}}^{1/2}\mathbf{a} + \kappa(\beta_0)\mathbf{v}^T\Sigma_{\mathbf{y}}^{1/2}\mathbf{a} + \tau(1 - \mathbf{a}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}})) \\ &= \begin{cases} \tau & \text{if } \tau\overline{\mathbf{x}} - \lambda\Sigma_{\mathbf{x}}^{1/2}\mathbf{u} = \tau\overline{\mathbf{y}} + \kappa(\beta_0)\Sigma_{\mathbf{y}}^{1/2}\mathbf{v} \\ -\infty & \text{otherwise} \end{cases} . \end{aligned}$$

Thus, the dual problem can further be changed to:

$$\max_{\tau,\mathbf{u},\mathbf{v}} \quad \tau : \|\mathbf{u}\|_2 \leq 1, \|\mathbf{v}\|_2 \leq 1, \tau\overline{\mathbf{x}} - \lambda\Sigma_{\mathbf{x}}^{1/2}\mathbf{u} = \tau\overline{\mathbf{y}} + \kappa(\beta_0)\Sigma_{\mathbf{y}}^{1/2}\mathbf{v} \ .$$

By defining $\ell := 1/\tau$, we rewrite the dual problem as

$$\min_{\ell,\mathbf{u},\mathbf{v}} \quad \ell : \overline{\mathbf{x}} - \lambda\Sigma_{\mathbf{x}}^{1/2}\mathbf{u} = \overline{\mathbf{y}} + \kappa(\beta_0)\Sigma_{\mathbf{y}}^{1/2}\mathbf{v}, \|\mathbf{u}\|_2 \leq \ell, \|\mathbf{v}\|_2 \leq \ell \ . \tag{20}$$

When the optimum is attained, we have

$$\tau_* = \lambda\|\Sigma_{\mathbf{x}}^{1/2}\mathbf{a}_*\|_2 + \kappa(\beta_0)\|\Sigma_{\mathbf{y}}^{1/2}\mathbf{a}_*\|_2 = 1/\ell_* \ .$$

We consider each side of (20) as an ellipsoid centered at the mean $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ and shaped by the weighted covariance matrices $\lambda\Sigma_{\mathbf{x}}$ and $\kappa(\beta_0)\Sigma_{\mathbf{y}}$ respectively:

$$\mathcal{H}_{\mathbf{x}}(\ell) = \{\mathbf{x} = \overline{\mathbf{x}} + \lambda\Sigma_{\mathbf{x}}^{1/2}\mathbf{u} : \|\mathbf{u}\|_2 \leq \ell\}, \ \mathcal{H}_{\mathbf{y}}(\ell) = \{\mathbf{y} = \overline{\mathbf{y}} + \kappa(\beta_0)\Sigma_{\mathbf{y}}^{1/2}\mathbf{v} : \|\mathbf{v}\|_2 \leq \ell\} \ .$$

The above optimization involves finding a minimum ℓ for which two ellipsoids intersect. For the optimum ℓ, these two ellipsoids are tangential to each other. We further note that, according to Lemma 3, at the optimum, λ_*, which is maximized via a series of the above procedures, satisfies

$$\begin{aligned} &1 = \lambda_*\|\Sigma_{\mathbf{x}}^{1/2}\mathbf{a}_*\|_2 + \kappa(\beta_0)\|\Sigma_{\mathbf{y}}^{1/2}\mathbf{a}_*\|_2 = \tau_* = 1/\ell_* \\ &\Rightarrow \ell_* = 1 \ . \end{aligned}$$

This means that the ellipsoid for the class $\mathbf{y}$ finally changes to the one centered at $\overline{\mathbf{y}}$, whose Mahalanobis distance to $\overline{\mathbf{y}}$ is exactly equal to $\kappa(\beta_0)$. Moreover, the ellipsoid for the class $\mathbf{x}$ is the one centered at $\overline{\mathbf{x}}$ and tangential to the ellipsoid for the class $\mathbf{y}$. In comparison, for MPM, two ellipsoids

grow with the same speed (with the same $\kappa(\alpha)$ and $\kappa(\beta)$). On the other hand, since MEMPM corresponds to solving a sequence of BMPMs, it similarly leads to a hyperplane tangential to two ellipsoids, which achieves to minimize the maximum of the worst-case Bayes error. Moreover, it is not necessarily attained in a balanced way as in MPM, i.e., two ellipsoids do not necessarily grow with the same speed and hence probably contain the unequal Mahalanobis distance from their corresponding centers. This is illustrated in Figure 3.

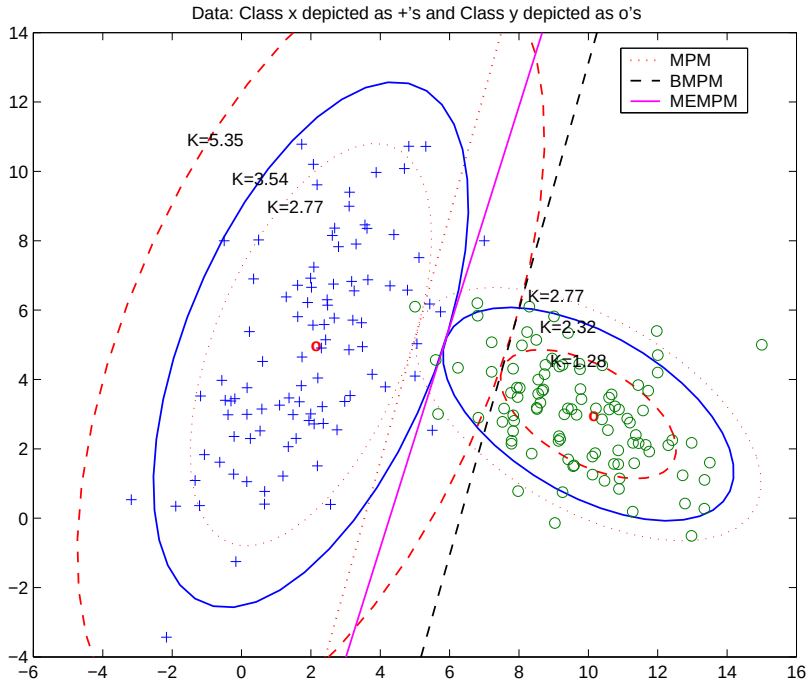


Figure 3: The geometrical interpretation of MEMPM and BMPM. Finding the optimal BMPM hyperplane corresponds to finding the decision plane (the black dashed line) tangential to an ellipsoid (the inner red dashed ellipsoid on the $\mathbf{y}$ side) , which is centered at $\overline{\mathbf{y}}$, shaped by the covariance $\Sigma_{\mathbf{y}}$ and whose Mahalanobis distance to $\overline{\mathbf{y}}$ is exactly equal to $\kappa(\beta_0)$ ($\kappa(\beta_0) = 1.28$ in this example). The worst-case accuracy α for $\mathbf{x}$ is determined by the Mahalanobis distance κ ($\kappa = 5.35$ in this example), at which, an ellipsoid (centered at $\overline{\mathbf{x}}$ and shaped by $\Sigma_{\mathbf{x}}$) is tangential to that $\kappa(\beta_0)$ ellipsoid, i.e., the outer red dashed ellipsoid on the $\mathbf{x}$ side. In comparison, MPM tries to find out the minimum equality-constrained κ, at which two ellipsoids for $\mathbf{x}$ and $\mathbf{y}$ intersect (both dotted red ellipsoids with $\kappa = 2.77$). For MEMPM, it achieves a tangent hyperplane in a non-balanced fashion, i.e., two ellipsoids may not attain the same κ but is globally optimal in the worst-case setting (see the solid blue ellipsoids).

3. Robust Version

In the above, the estimates of means and covariance matrices are assumed reliable. We now consider how the probabilistic framework in (1) changes in the face of variation of the means and covariance matrices:

$$\begin{aligned}
&\max_{\alpha,\beta,\mathbf{a}\neq\mathbf{0},b} \quad \theta\alpha+(1-\theta)\beta \quad \text{s.t.}\\
&\inf_{\mathbf{x}\sim(\bar{\mathbf{x}},\Sigma_{\mathbf{x}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x}\geq b\}\geq\alpha, \forall(\bar{\mathbf{x}},\Sigma_{\mathbf{x}})\in\mathcal{X}\ ,\\
&\inf_{\mathbf{y}\sim(\bar{\mathbf{y}},\Sigma_{\mathbf{y}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y}\leq b\}\geq\beta, \forall(\bar{\mathbf{y}},\Sigma_{\mathbf{y}})\in\mathcal{Y}\ ,
\end{aligned}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the sets of means and covariance matrices and are the subsets of $\mathbb{R}\times\mathcal{P}_n^+$, where $\mathcal{P}_n^+$ is the set of $n\times n$ symmetric positive semidefinite matrices.

Motivated by the tractability of the problem and from a statistical viewpoint, a specific setting of $\mathcal{X}$ and $\mathcal{Y}$ has been proposed in Lanckriet et al. (2002b). However, these authors consider the same variations of the means for two classes, which is easy to handle but less general. Now, considering the unequal treatment of each class, we propose the following setting, which is in a more general and complete form:

$$\begin{aligned}
\mathcal{X} &= \left\{(\bar{\mathbf{x}},\Sigma_{\mathbf{x}})\,|\,(\bar{\mathbf{x}}-\bar{\mathbf{x}}^0){\Sigma_{\mathbf{x}}}^{-1}(\bar{\mathbf{x}}-\bar{\mathbf{x}}^0)\leq\nu_{\mathbf{x}}^2,\ \|\Sigma_{\mathbf{x}}-{\Sigma_{\mathbf{x}}}^0\|_F\leq\rho_{\mathbf{x}}\right\},\\
\mathcal{Y} &= \left\{(\bar{\mathbf{y}},\Sigma_{\mathbf{y}})\,|\,(\bar{\mathbf{y}}-\bar{\mathbf{y}}^0){\Sigma_{\mathbf{y}}}^{-1}(\bar{\mathbf{y}}-\bar{\mathbf{y}}^0)\leq\nu_{\mathbf{y}}^2,\ \|\Sigma_{\mathbf{y}}-{\Sigma_{\mathbf{y}}}^0\|_F\leq\rho_{\mathbf{y}}\right\},
\end{aligned}$$

where $\bar{\mathbf{x}}^0$, $\Sigma_{\mathbf{x}}^0$ are the "nominal" mean and covariance matrices obtained through estimation. Parameters $\nu_{\mathbf{x}}$, $\nu_{\mathbf{y}}$, $\rho_{\mathbf{x}}$, and $\rho_{\mathbf{y}}$ are positive constants. The matrix norm is defined as the Frobenius norm: $\|M\|_F^2=\mathrm{Tr}(M^TM)$.

With the equality assumption for the variations of the means for two classes, the parameters $\nu_{\mathbf{x}}$ and $\nu_{\mathbf{y}}$ are required equal in Lanckriet et al. (2002b). This enables the direct usage of the MPM optimization in its robust version. However, the assumption may not be valid in real cases. Moreover, in MEMPM, the assumption is also unnecessary and inappropriate. This will be demonstrated later in the experiment.

By applying the results from Lanckriet et al. (2002b), we obtain the robust MEMPM as

$$\begin{aligned}
&\max_{\alpha,\beta,\mathbf{a}\neq\mathbf{0},b} \quad \theta\alpha+(1-\theta)\beta \quad \text{s.t.}\\
&-b+\mathbf{a}^T\bar{\mathbf{x}}^0\geq(\kappa(\alpha)+\nu_{\mathbf{x}})\sqrt{\mathbf{a}^T({\Sigma_{\mathbf{x}}}^0+\rho_{\mathbf{x}}I_n)\mathbf{a}},\\
&b-\mathbf{a}^T\bar{\mathbf{y}}^0\geq(\kappa(\beta)+\nu_{\mathbf{y}})\sqrt{\mathbf{a}^T({\Sigma_{\mathbf{y}}}^0+\rho_{\mathbf{y}}I_n)\mathbf{a}}.
\end{aligned}$$

Analogously, we transform the above optimization problem to

$$\max_{\alpha,\beta,\mathbf{a}\neq\mathbf{0}} \theta\frac{\kappa_r^2(\alpha)}{1+\kappa_r^2(\alpha)}+(1-\theta)\beta \quad \text{s.t. } \mathbf{a}^T(\bar{\mathbf{x}}^0-\bar{\mathbf{y}}^0)=1,$$

where $\kappa_r(\alpha)=\max\left(\frac{1-(\kappa(\beta)+\nu_{\mathbf{y}})\sqrt{\mathbf{a}^T({\Sigma_{\mathbf{y}}}^0+\rho_{\mathbf{y}}I_n)\mathbf{a}}}{\sqrt{\mathbf{a}^T({\Sigma_{\mathbf{x}}}^0+\rho_{\mathbf{x}}I_n)\mathbf{a}}}-\nu_{\mathbf{x}},0\right)$ and thus can be solved by the SBMPM method. The optimal b is therefore calculated by

$$\begin{aligned}
b_* &= {\mathbf{a}_*}^T\bar{\mathbf{x}}^0-(\kappa(\alpha_*)+\nu_{\mathbf{x}})\sqrt{{\mathbf{a}_*}^T({\Sigma_{\mathbf{x}}}^0+\rho_{\mathbf{x}}I_n)\mathbf{a}_*}\\
&= {\mathbf{a}_*}^T\bar{\mathbf{y}}^0+(\kappa(\beta_*)+\nu_{\mathbf{y}})\sqrt{{\mathbf{a}_*}^T({\Sigma_{\mathbf{y}}}^0+\rho_{\mathbf{y}}I_n)\mathbf{a}_*}.
\end{aligned}$$

Remarks. Interestingly, if MPM is treated with unequal robust parameters $\nu_{\mathbf{x}}$ and $\nu_{\mathbf{y}}$, it leads to solving an optimization similar to MEMPM, since $\kappa(\alpha)+\nu_{\mathbf{x}}$ will not be equal to $\kappa(\alpha)+\nu_{\mathbf{y}}$. In addition, similar to the robust MPM, when applied in practice, the specific values of $\nu_{\mathbf{x}}$, $\nu_{\mathbf{y}}$, $\rho_{\mathbf{x}}$, and $\rho_{\mathbf{y}}$ can be provided based on the central limit theorem or the resampling method.

4. Kernelization

We note that, in the above, the classifier derived from MEMPM is given in a linear configuration. In order to handle nonlinear classification problems, in this section, we seek to use the kernelization trick (Schölkopf and Smola, 2002) to map the *n*-dimensional data points into a high-dimensional feature space $\mathbb{R}^f$, where a linear classifier corresponds to a nonlinear hyperplane in the original space.

Since the optimization of MEMPM corresponds to a sequence of BMPM optimization problems, this model will naturally inherit the kernelization ability of BMPM. We thus in the following mainly address the kernelization of BMPM.

Assuming training data points are represented by $\{\mathbf{x}_i\}_{i=1}^{N_{\mathbf{x}}}$ and $\{\mathbf{y}_j\}_{j=1}^{N_{\mathbf{y}}}$ for the class $\mathbf{x}$ and class $\mathbf{y}$, respectively, the kernel mapping can be formulated as

$$\mathbf{x} \to \varphi(\mathbf{x}) \sim (\overline{\varphi(\mathbf{x})}, \Sigma_{\varphi(\mathbf{x})}),$$
$$\mathbf{y} \to \varphi(\mathbf{y}) \sim (\overline{\varphi(\mathbf{y})}, \Sigma_{\varphi(\mathbf{y})}),$$

where $\varphi : \mathbb{R}^n \to \mathbb{R}^f$ is a mapping function. The corresponding linear classifier in $\mathbb{R}^f$ is $\mathbf{a}^T\varphi(\mathbf{z}) = b$, where $\mathbf{a}$, $\varphi(\mathbf{z}) \in \mathbb{R}^f$, and $b \in \mathbb{R}$. Similarly, the transformed FP optimization in BMPM can be written as

$$\max_{\mathbf{a}} \frac{1-\kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_{\varphi(\mathbf{y})}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_{\varphi(\mathbf{x})}\mathbf{a}}} \quad \text{s.t.} \quad \mathbf{a}^T(\overline{\varphi(\mathbf{x})} - \overline{\varphi(\mathbf{y})}) = 1. \tag{21}$$

However, to make the kernel work, we need to represent the final decision hyperplane and the optimization in a kernel form, $K(\mathbf{z}_1,\mathbf{z}_2) = \varphi(\mathbf{z}_1)^T\varphi(\mathbf{z}_2)$, namely an inner product form of the mapping data points.

4.1 Kernelization Theory for BMPM

In the following, we demonstrate that, although BMPM possesses a significantly different optimization form from MPM, the kernelization theory proposed in Lanckriet et al. (2002b) is still viable, provided that suitable estimates for means and covariance matrices are applied therein.

We first state a theory similar to Corollary 5 of Lanckriet et al. (2002b) and prove its validity in BMPM.

Corollary 7 *If the estimates of means and covariance matrices are given in BMPM as*

$$\overline{\varphi(\mathbf{x})} = \sum_{i=1}^{N_{\mathbf{x}}} \lambda_i \varphi(\mathbf{x}_i), \quad \overline{\varphi(\mathbf{y})} = \sum_{j=1}^{N_{\mathbf{y}}} \omega_j \varphi(\mathbf{y}_j),$$

$$\Sigma_{\varphi(\mathbf{x})} = \rho_{\mathbf{x}}\mathbf{I}_n + \sum_{i=1}^{N_{\mathbf{x}}} \Lambda_i(\varphi(\mathbf{x}_i) - \overline{\varphi(\mathbf{x})})(\varphi(\mathbf{x}_i) - \overline{\varphi(\mathbf{x})})^T ,$$

$$\Sigma_{\varphi(\mathbf{y})} = \rho_{\mathbf{y}}\mathbf{I}_n + \sum_{j=1}^{N_{\mathbf{y}}} \Omega_j(\varphi(\mathbf{y}_j) - \overline{\varphi(\mathbf{y})})(\varphi(\mathbf{y}_j) - \overline{\varphi(\mathbf{y})})^T ,$$

where $\mathbf{I}_n$ *is the identity matrix of dimension n, then the optimal* $\mathbf{a}$ *in problem (21) lies in the space spanned by the training points.*

Proof Similar to Lanckriet et al. (2002b), we write $\mathbf{a} = \mathbf{a}_p + \mathbf{a}_d$, where $\mathbf{a}_p$ is the projection of $\mathbf{a}$ in the vector space spanned by all the training data points and $\mathbf{a}_d$ is the orthogonal component to this span space. It can be easily verified that (21) changes to maximize the following:

$$\frac{1-\kappa(\beta_0)\sqrt{\mathbf{a}_p^T \sum_{i=1}^{N_{\mathbf{x}}} \Lambda_i(\varphi(\mathbf{x}_i) - \overline{\varphi(\mathbf{x})})(\varphi(\mathbf{x}_i) - \overline{\varphi(\mathbf{x})})^T \mathbf{w}_p + \rho_{\mathbf{x}}(\mathbf{a}_p^T\mathbf{a}_p + \mathbf{w}_d^T\mathbf{a}_d)}}{\sqrt{\mathbf{a}_p^T \sum_{j=1}^{N_{\mathbf{y}}} \Omega_j(\varphi(\mathbf{y}_j) - \overline{\varphi(\mathbf{y})})(\varphi(\mathbf{y}_j) - \overline{\varphi(\mathbf{y})})^T \mathbf{a}_p + \rho_{\mathbf{y}}(\mathbf{a}_p^T\mathbf{a}_p + \mathbf{a}_d^T\mathbf{a}_d)}}$$

subject to the constraints of $\mathbf{a}_p^T(\overline{\varphi(\mathbf{x})} - \overline{\varphi(\mathbf{y})}) = 1$.

Since we intend to maximize the fractional form and both the denominator and the numerator are positive, the denominator needs to be as small as possible and the numerator needs to be as large as possible. This would finally lead to $\mathbf{a}_d = \mathbf{0}$. In other words, the optimal $\mathbf{a}$ lies in the vector space spanned by all the training data points. Note that the introduction of $\rho_{\mathbf{x}}$ and $\rho_{\mathbf{y}}$ enables a direct application of the robust estimates in the kernelization. ■

According to Corollary 7, if appropriate estimates of means and covariance matrices are applied, the optimal $\mathbf{a}$ can be written as the linear combination of training points. In particular, if we obtain the means and covariance matrices as the plug-in estimates, i.e.,

$$\overline{\varphi(\mathbf{x})} = \frac{1}{N_{\mathbf{x}}} \sum_{i=1}^{N_{\mathbf{x}}} \varphi(\mathbf{x}_i) ,$$

$$\overline{\varphi(\mathbf{y})} = \frac{1}{N_{\mathbf{y}}} \sum_{j=1}^{N_{\mathbf{y}}} \varphi(\mathbf{y}_j) ,$$

$$\Sigma_{\varphi(\mathbf{x})} = \frac{1}{N_{\mathbf{x}}} \sum_{i=1}^{N_{\mathbf{x}}} (\varphi(\mathbf{x}_i) - \overline{\varphi(\mathbf{x})})(\varphi(\mathbf{x}_i) - \overline{\varphi(\mathbf{x})})^T ,$$

$$\Sigma_{\varphi(\mathbf{y})} = \frac{1}{N_{\mathbf{y}}} \sum_{j=1}^{N_{\mathbf{y}}} (\varphi(\mathbf{y}_j) - \overline{\varphi(\mathbf{y})})(\varphi(\mathbf{y}_j) - \overline{\varphi(\mathbf{y})})^T ,$$

we can write $\mathbf{a}$ as

$$\mathbf{a} = \sum_{i=1}^{N_{\mathbf{x}}} \mu_i \varphi(\mathbf{x}_i) + \sum_{j=1}^{N_{\mathbf{y}}} \upsilon_j \varphi(\mathbf{y}_j), \tag{22}$$

where the coefficients $\mu_i, \upsilon_j \in \mathbb{R}, i = 1,\ldots,N_{\mathbf{x}}, j = 1,\ldots,N_{\mathbf{y}}$.

By simply substituting (22) and four plug-in estimates into (21), we can obtain the Kernelization Theorem of BMPM.

Kernelization Theorem of BMPM *The optimal decision hyperplane of the problem (21) involves solving the Fractional Programming problem*

$$\kappa(\alpha_*) = \max_{\mathbf{w}\neq\mathbf{0}} \frac{1-\kappa(\beta_0)\sqrt{\frac{1}{N_\mathbf{y}}\mathbf{w}^T\tilde{\mathbf{K}}_\mathbf{y}^T\tilde{\mathbf{K}}_\mathbf{y}\mathbf{w}}}{\sqrt{\frac{1}{N_\mathbf{x}}\mathbf{w}^T\tilde{\mathbf{K}}_\mathbf{x}^T\tilde{\mathbf{K}}_\mathbf{x}\mathbf{w}}} \quad s.t. \quad \mathbf{w}^T(\tilde{\mathbf{k}}_\mathbf{x}-\tilde{\mathbf{k}}_\mathbf{y}) = 1\,. \tag{23}$$

The intercept b is calculated as

$$b_* = \mathbf{w}_*^T\tilde{\mathbf{k}}_\mathbf{x} - \kappa(\alpha_*)\sqrt{\frac{1}{N_\mathbf{x}}\mathbf{w}_*^T\tilde{\mathbf{K}}_\mathbf{x}^T\tilde{\mathbf{K}}_\mathbf{x}\mathbf{w}_*} = \mathbf{w}_*^T\tilde{\mathbf{k}}_\mathbf{y} + \kappa(\beta_0)\sqrt{\frac{1}{N_\mathbf{y}}\mathbf{w}_*^T\tilde{\mathbf{K}}_\mathbf{y}^T\tilde{\mathbf{K}}_\mathbf{y}\mathbf{w}_*}\,,$$

where $\kappa(\alpha_*)$ is obtained when (23) attains its optimum $(\mathbf{w}_*, b_*)$. For the robust version of BMPM, we can incorporate the variations of the means and covariances by conducting the following replacements:

$$\begin{aligned}
&\frac{1}{N_\mathbf{x}}\mathbf{w}_*^T\tilde{\mathbf{K}}_\mathbf{x}^T\tilde{\mathbf{K}}_\mathbf{x}\mathbf{w}_* \to \mathbf{w}_*^T(\frac{1}{N_\mathbf{x}}\tilde{\mathbf{K}}_\mathbf{x}^T\tilde{\mathbf{K}}_\mathbf{x} + \rho_\mathbf{x}\mathbf{K})\mathbf{w}_*\,,\\
&\frac{1}{N_\mathbf{y}}\mathbf{w}_*^T\tilde{\mathbf{K}}_\mathbf{y}^T\tilde{\mathbf{K}}_\mathbf{y}\mathbf{w}_* \to \mathbf{w}_*^T(\frac{1}{N_\mathbf{y}}\tilde{\mathbf{K}}_\mathbf{y}^T\tilde{\mathbf{K}}_\mathbf{y} + \rho_\mathbf{y}\mathbf{K})\mathbf{w}_*\,,\\
&\kappa(\beta_0) \to \kappa(\beta_0) + \mu_\mathbf{y}\,,\\
&\kappa(\alpha_*) \to \kappa(\alpha_*) + \mu_\mathbf{x}\,.
\end{aligned}$$

The optimal decision hyperplane can be represented as a linear form in the kernel space

$$f(\mathbf{z}) = \sum_{i=1}^{N_\mathbf{x}} \mathbf{w}_{*i}\mathbf{K}(\mathbf{z},\mathbf{x}_i) + \sum_{i=1}^{N_\mathbf{y}} \mathbf{w}_{*N_\mathbf{x}+i}\mathbf{K}(\mathbf{z},\mathbf{y}_i) - b_*.$$

The notation in the above are defined in Table 1.

5. Experiments

In this section, we first evaluate our model on a synthetic data set. Then we compare the performance of MEMPM with that of MPM, on six real world benchmark data sets. To demonstrate that BMPM is ideal for imposing a specified bias in classification, we also implement it on the Heart-disease data set. The means and covariance matrices for two classes are obtained directly from the training data sets by plug-in estimations. The prior probability θ is given by the proportion of $\mathbf{x}$ data in the training set.

5.1 Model Illustration on a Synthetic Data Set

To verify that the MEMPM model achieves the minimum Bayes error rate in the Gaussian distribution, we synthetically generate two classes of two-dimensional Gaussian data. As plotted in Figure 4(a), data associated with the class $\mathbf{x}$ are generated with the mean $\overline{\mathbf{x}}$ as $[3,0]^T$ and the covariance matrix $\Sigma_\mathbf{x}$ as $[4,\ 0; 0,\ 1]$, while data associated with the class $\mathbf{y}$ are generated with the mean $\overline{\mathbf{y}}$ as $[-1,0]^T$ and the covariance matrix $\Sigma_\mathbf{y}$ as $[1,\ 0; 0,\ 5]$. The solved decision hyperplane $Z_1 = 0.333$

Notation		
$\mathbf{z} \in \mathbb{R}^{N_\mathbf{x}+N_\mathbf{y}}$	$\mathbf{z}_i :=$	$\mathbf{x}_i \quad i = 1,2,\ldots,N_\mathbf{x}$.
	$\mathbf{z}_i :=$	$\mathbf{y}_{i-N_\mathbf{x}} \quad i = N_\mathbf{x}+1, N_\mathbf{x}+2, \ldots, N_\mathbf{x}+N_\mathbf{y}$.
$\mathbf{w} \in \mathbb{R}^{N_\mathbf{x}+N_\mathbf{y}}$	$\mathbf{w} :=$	$[\mu_1,\ldots,\mu_{N_\mathbf{x}},\upsilon_1,\ldots,\upsilon_{N_\mathbf{y}}]^T$.
$\mathbf{K}$ is Gram matrix	$\mathbf{K}_{i,j} :=$	$\varphi(\mathbf{z}_i)^T\varphi(\mathbf{z}_j)$.
	$\mathbf{K}_\mathbf{x} :=$	$\begin{pmatrix} \mathbf{K}_{1,1} & \mathbf{K}_{1,2} & \cdots & \mathbf{K}_{1,N_\mathbf{x}+N_\mathbf{y}} \\ \mathbf{K}_{2,1} & \mathbf{K}_{2,2} & \cdots & \mathbf{K}_{2,N_\mathbf{x}+N_\mathbf{y}} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{K}_{N_\mathbf{x},1} & \mathbf{K}_{N_\mathbf{x},2} & \cdots & \mathbf{K}_{N_\mathbf{x},N_\mathbf{x}+N_\mathbf{y}} \end{pmatrix}$.
	$\mathbf{K}_\mathbf{y} :=$	$\begin{pmatrix} \mathbf{K}_{N_\mathbf{x}+1,1} & \mathbf{K}_{N_\mathbf{x}+1,2} & \cdots & \mathbf{K}_{N_\mathbf{x}+1,N_\mathbf{x}+N_\mathbf{y}} \\ \mathbf{K}_{N_\mathbf{x}+2,1} & \mathbf{K}_{N_\mathbf{x}+2,2} & \cdots & \mathbf{K}_{N_\mathbf{x}+2,N_\mathbf{x}+N_\mathbf{y}} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{K}_{N_\mathbf{x}+N_\mathbf{y},1} & \mathbf{K}_{N_\mathbf{x}+N_\mathbf{y},2} & \cdots & \mathbf{K}_{N_\mathbf{x}+N_\mathbf{y},N_\mathbf{x}+N_\mathbf{y}} \end{pmatrix}$.
$\tilde{\mathbf{k}}_\mathbf{x}, \tilde{\mathbf{k}}_\mathbf{y} \in \mathbb{R}^{N_\mathbf{x}+N_\mathbf{y}}$	$[\tilde{\mathbf{k}}_\mathbf{x}]_i :=$	$\frac{1}{N_\mathbf{x}}\sum_{j=1}^{N_\mathbf{x}} \mathbf{K}(\mathbf{x}_j,\mathbf{z}_i)$.
	$[\tilde{\mathbf{k}}_\mathbf{y}]_i :=$	$\frac{1}{N_\mathbf{y}}\sum_{j=1}^{N_\mathbf{y}} \mathbf{K}(\mathbf{y}_j,\mathbf{z}_i)$.
$\mathbf{1}_{N_\mathbf{x}} \in \mathbb{R}^{N_\mathbf{x}}$	$\mathbf{1}_i :=$	$1 \quad i = 1,2,\ldots N_\mathbf{x}$.
$\mathbf{1}_{N_\mathbf{y}} \in \mathbb{R}^{N_\mathbf{y}}$	$\mathbf{1}_i :=$	$1 \quad i = 1,2,\ldots N_\mathbf{y}$.
$\tilde{\mathbf{K}} :=$	$\begin{pmatrix} \tilde{\mathbf{K}}_\mathbf{x} \\ \tilde{\mathbf{K}}_\mathbf{y} \end{pmatrix} :=$	$\begin{pmatrix} \mathbf{K}_\mathbf{x} - \mathbf{1}_{N_\mathbf{x}}\tilde{\mathbf{k}}_\mathbf{x}^T \\ \mathbf{K}_\mathbf{y} - \mathbf{1}_{N_\mathbf{y}}\tilde{\mathbf{k}}_\mathbf{y}^T \end{pmatrix}$.

Table 1: Notation used in Kernelization Theorem of BMPM

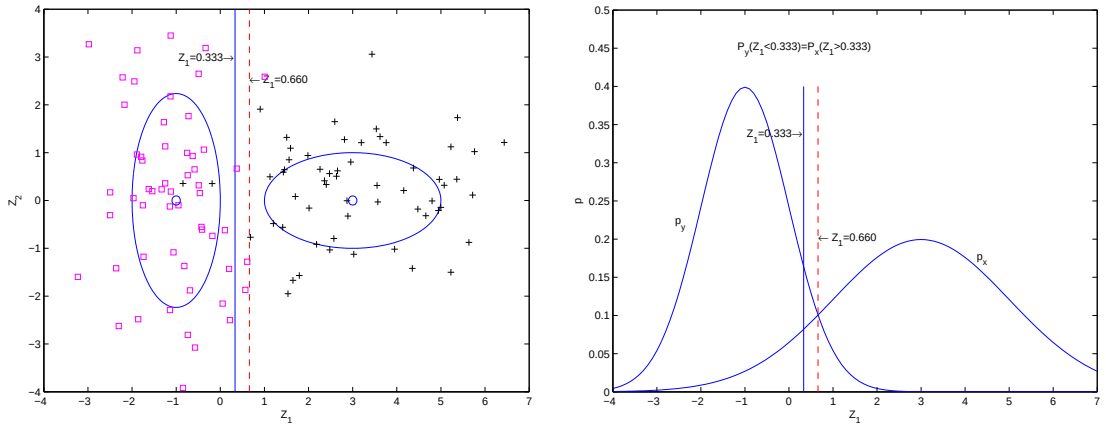
given by MPM is plotted as the solid blue line and the solved decision hyperplane $Z_1 = 0.660$ given by MEMPM is plotted as the dashed red line. From the geometrical interpretation, both hyperplanes should be perpendicular to the Z_1 axis.

As shown in Figure 4(b), the MEMPM hyperplane exactly represents the optimal thresholding under the distributions of the first dimension for two classes of data, i.e., the intersection point of two density functions. On the other hand, we find that, the MPM hyperplane exactly corresponds to the thresholding point with the same error rate for two classes of data, since the cumulative distributions $P_\mathbf{x}(Z_1 < 0.333)$ and $P_\mathbf{y}(Z_1 > 0.333)$ are exactly the same.

5.2 Evaluations on Benchmark Data Sets

We next evaluate our algorithm on six benchmark data sets. Data for Twonorm problem were generated according to Breiman (1997). The remaining five data sets (Breast, Ionosphere, Pima, Heart-disease, and Vote) were obtained from the UCI machine learning repository (Blake and Merz, 1998). Since handling the missing attribute values is out of the scope of this paper, we simply remove instances with missing attribute values in these data sets.

We randomly partition data into 90% training and 10% test sets. The final results are averaged over 50 random partitions of data. We compare the performance of MEMPM and MPM in both the linear setting and Gaussian kernel setting. The width parameter (σ) for the Gaussian kernel is obtained via cross validations over 50 random partitions of the training set. The experimental results are summarized in Table 2 and Table 3 for the linear kernel and Gaussian kernel respectively.



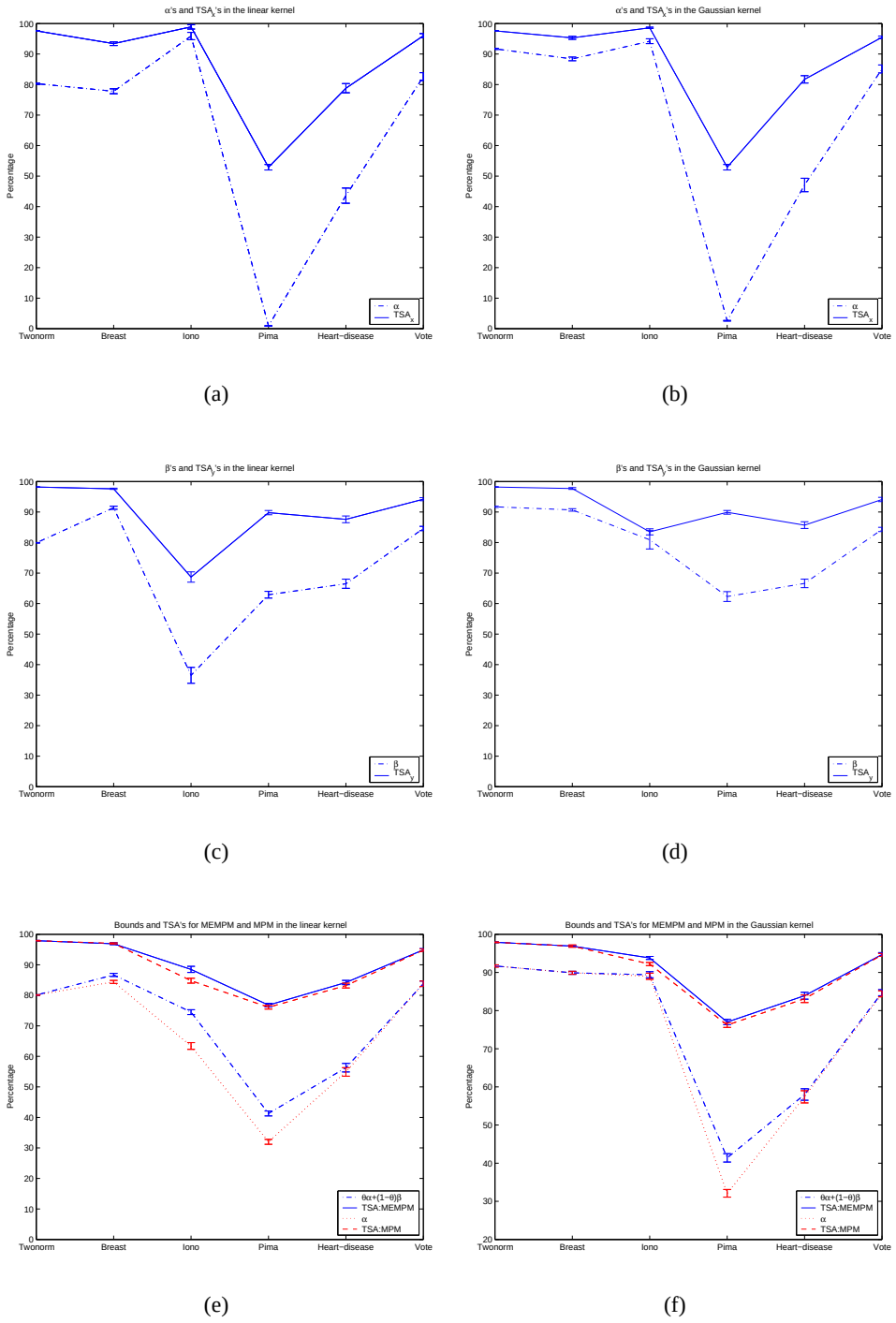
(a) (b)

Figure 4: An experiment on a synthetic data set. The decision hyperplane derived from MEMPM (the dashed red line) exactly corresponds to the optimal thresholding point, i.e., the intersection point, while the decision hyperplane given by MPM (the solid blue line) corresponds to the point in which the error rates for the two classes of data are equal.

From the results, we can see that our MEMPM demonstrates better performance than MPM in both the linear setting and Gaussian kernel setting. Moreover, in these benchmark data sets, the MEMPM hyperplanes are obtained with significantly unequal α and β except in Twonorm. This further confirms the validity of our proposition, i.e., the optimal minimax machine is not certain to achieve the same worst-case accuracies for two classes. Twonorm is not an exception to this. The two classes of data in Twonorm are generated under the multivariate normal distributions with the same covariance matrices. In this special case, the intersection point of two density functions will exactly represent the optimal thresholding point and the one with the same error rate for each class as well. Another important finding is that the accuracy bounds, namely $\theta\alpha+(1-\theta)\beta$ in MEMPM and α in MPM, are all increased in the Gaussian kernel setting when compared with those in the linear setting. This shows the advantage of the kernelized probability machine over the linear probability machine.

In addition, to show the relationship between the bounds and the test set accuracies (TSA) clearly, we plot them in Figure 5. As observed, the test set accuracies including $\text{TSA}_{\mathbf{x}}$ (for class $\mathbf{x}$), $\text{TSA}_{\mathbf{y}}$ (for the class $\mathbf{y}$), and the overall accuracies TSA are all greater than their corresponding accuracy bounds in both MPM and MEMPM. This demonstrates how the accuracy bound can serve as the performance indicator on future data. It is also observed that the overall worst-case accuracies $\theta\alpha+(1-\theta)\beta$ in MEMPM are greater than α in MPM both in the linear and Gaussian setting. This again demonstrates the superiority of MEMPM to MPM.

Since the lower bounds keep well within the test accuracies in the above experimental results, we do not perform the robust version of both models for the real world data sets. To see how the robust version works, we generate two classes of Gaussian data. As illustrated in Figure 6, $\mathbf{x}$ data are sampled from the Gaussian distribution with the mean as $[3,0]^T$ and the covariance as



(a) (b)

(c) (d)

(e) (f)

Figure 5: Bounds and test set accuracies. The test accuracies including $\text{TSA}_{\mathbf{x}}$ (for the class **x**), $\text{TSA}_{\mathbf{y}}$ (for the class **y**), and the overall accuracies TSA are all greater than their corresponding accuracy bounds in both MPM and MEMPM. This demonstrates how the accuracy bound can serve as the performance indicator on future data.

Data Set	MEMPM				MPM	
	α	β	$\theta\alpha+(1-\theta)\beta$	Accuracy	α	Accuracy
Twonorm(%)	$80.3\pm0.2\%$	$79.9\pm0.1\%$	$80.1\pm0.1\%$	$\mathbf{97.9\pm0.1}\%$	$80.1\pm0.1\%$	$\mathbf{97.9\pm0.1}\%$
Breast(%)	$77.8\pm0.8\%$	$91.4\pm0.5\%$	$86.7\pm0.5\%$	$96.9\pm0.3\%$	$84.4\pm0.5\%$	$\mathbf{97.0\pm0.2}\%$
Ionosphere(%)	$95.9\pm1.2\%$	$36.5\pm2.6\%$	$74.5\pm0.8\%$	$\mathbf{88.5\pm1.0}\%$	$63.4\pm1.1\%$	$84.8\pm0.8\%$
Pima(%)	$0.9\pm0.0\%$	$62.9\pm1.1\%$	$41.3\pm0.8\%$	$\mathbf{76.8\pm0.6}\%$	$32.0\pm0.8\%$	$76.1\pm0.6\%$
Heart-disease(%)	$43.6\pm2.5\%$	$66.5\pm1.5\%$	$56.3\pm1.4\%$	$\mathbf{84.2\pm0.7}\%$	$54.9\pm1.4\%$	$83.2\pm0.8\%$
Vote(%)	$82.6\pm1.3\%$	$84.6\pm0.7\%$	$83.9\pm0.9\%$	$\mathbf{94.9\pm0.4}\%$	$83.8\pm0.9\%$	$94.8\pm0.4\%$

Table 2: Lower bound α, β, and test accuracy compared to MPM in the linear setting.

Data Set	MEMPM				MPM	
	α	β	$\theta\alpha+(1-\theta)\beta$	Accuracy	α	Accuracy
Twonorm(%)	$91.7\pm0.2\%$	$91.7\pm0.2\%$	$91.7\pm0.2\%$	$\mathbf{97.9\pm0.1}\%$	$91.7\pm0.2\%$	$\mathbf{97.9\pm0.1}\%$
Breast(%)	$88.4\pm0.6\%$	$90.7\pm0.4\%$	$89.9\pm0.4\%$	$\mathbf{96.9\pm0.2}\%$	$89.9\pm0.4\%$	$\mathbf{96.9\pm0.3}\%$
Ionosphere(%)	$94.2\pm0.8\%$	$80.9\pm3.0\%$	$89.4\pm0.8\%$	$\mathbf{93.8\pm0.4}\%$	$89.0\pm0.8\%$	$92.2\pm0.4\%$
Pima(%)	$2.6\pm0.1\%$	$62.3\pm1.6\%$	$41.4\pm1.1\%$	$\mathbf{77.0\pm0.7}\%$	$32.1\pm1.0\%$	$76.2\pm0.6\%$
Heart-disease(%)	$47.1\pm2.2\%$	$66.6\pm1.4\%$	$58.0\pm1.5\%$	$\mathbf{83.9\pm0.9}\%$	$57.4\pm1.6\%$	$83.1\pm1.0\%$
Vote(%)	$85.1\pm1.3\%$	$84.3\pm0.7\%$	$84.7\pm0.8\%$	$\mathbf{94.7\pm0.5}\%$	$84.4\pm0.8\%$	$94.6\pm0.4\%$

Table 3: Lower bound α, β, and test accuracy compared to MPM with the Gaussian kernel.

$[1\ 0;0\ 3]$, while $\mathbf{y}$ data are sampled from another Gaussian distribution with the mean as $[-3,0]^T$ and the covariance as $[3\ 0;0\ 1]$. We randomly select 10 points of each class for training and leave the remaining points for test from the above synthetic data set. We present the result below.

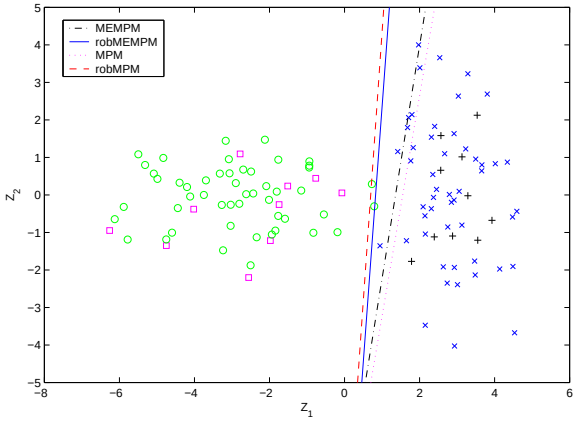
First, we calculate the corresponding means, $\bar{\mathbf{x}}^0$ and $\bar{\mathbf{y}}^0$ and covariance matrices, ${\Sigma_{\mathbf{x}}}^0$ and ${\Sigma_{\mathbf{y}}}^0$ and plug them into the linear MPM and the linear MEMPM. We obtain the MPM decision line (magenta dotted line) with a lower bound (assuming the Gaussian distribution) being 99.1% and the MEMPM decision line (black dash-dot line) with a lower bound of 99.7%. However, for the test set, we obtain the accuracies of only 93.0% for MPM and 97.0% for MEMPM (see Figure 6(a)). This obviously violates the lower bound.

Based on our knowledge of the real means and covariance matrices in this example, we set the parameters as

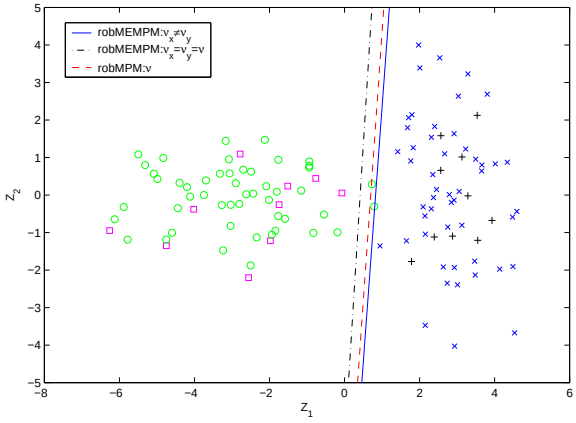
$$\nu_{\mathbf{x}} = \sqrt{(\bar{\mathbf{x}}-\bar{\mathbf{x}}^0)^T{\Sigma_{\mathbf{x}}}^{-1}(\bar{\mathbf{x}}-\bar{\mathbf{x}}^0)} = 0.046,\ \nu_{\mathbf{y}} = \sqrt{(\bar{\mathbf{y}}-\bar{\mathbf{y}}^0)^T{\Sigma_{\mathbf{y}}}^{-1}(\bar{\mathbf{y}}-\bar{\mathbf{y}}^0)} = 0.496,$$
$$\nu = \max(\nu_{\mathbf{x}},\nu_{\mathbf{y}}), \quad \rho_{\mathbf{x}} = \|\Sigma_{\mathbf{x}}-{\Sigma_{\mathbf{x}}}^0\|_F = 1.561, \quad \rho_{\mathbf{y}} = \|\Sigma_{\mathbf{y}}-{\Sigma_{\mathbf{y}}}^0\|_F = 0.972\ .$$

We then train the robust linear MPM and the robust linear MEMPM by these parameters and obtain the robust MPM decision line (red dashed line), and the robust MEMPM decision line (blue solid line), as seen in Figure 6(a). The lower bounds decrease to 87.3% for MPM and 93.2% for MEMPM respectively, but the test accuracies increase to 98.0% for MPM and 100.0% for MEMPM. Obviously, the lower bounds accord with the test accuracies.

Note that in the above, the robust MEMPM also achieves better performance than the robust MPM. Moreover, $\nu_{\mathbf{x}}$ and $\nu_{\mathbf{y}}$ are not necessarily the same. To see the result of MEMPM when $\nu_{\mathbf{x}}$ and $\nu_{\mathbf{y}}$ are forced to be the same, we train the robust MEMPM again by setting the parameters as $\nu_{\mathbf{x}} = \nu_{\mathbf{y}} = \nu$ as used in MPM. We obtain the corresponding decision line (black dash-dot line) as seen in Figure 6(b). The lower bound decreases to 91.0% and the test accuracy decreases to 98.0%. The above example indicates how the robust MEMPM clearly improves on the standard MEMPM when a bias is incorporated by inaccurate plug-in estimates and also validates that $\nu_{\mathbf{x}}$ need not be equal to $\nu_{\mathbf{y}}$.



(a)



(b)

Figure 6: An example in $\mathbb{R}^2$ demonstrates the results of robust versions of MEMPM and MPM. Training points are indicated with black $+$'s for the class $\mathbf{x}$ and magenta $\square$'s for the class $\mathbf{y}$. Test points are represented by blue $\times$'s for the class $\mathbf{x}$ and by green o's for the class $\mathbf{y}$. In (a), the robust MEMPM outperforms both MEMPM and the robust MPM. In (b), the robust MEMPM with $\nu_{\mathbf{x}} \neq \nu_{\mathbf{y}}$ outperforms the robust MEMPM with $\nu_{\mathbf{x}} = \nu_{\mathbf{y}}$.

5.3 Evaluations of BMPM on the Heart-Disease Data Set

To demonstrate the advantages of the BMPM model in dealing with biased classification, we implement BMPM on the Heart-disease data set, where a different treatment for different classes is necessary. The **x** class is associated with subjects with heart disease, whereas the **y** class corresponds to subjects without heart disease. Obviously, a bias should be considered for **x**, since misclassification of an **x** case into the opposite class would delay the therapy and may have a higher risk than the other way round. Similarly, we randomly partition data into 90% training and 10% test sets. Also, the width parameter (σ) for the Gaussian kernel is obtained via cross validations over 50 random partitions of the training set. We repeat the above procedures 50 times and report the average results.

By intentionally varying β_0 from 0 to 1, we obtain a series of test accuracies, including the **x** accuracy, $\text{TSA}_{\mathbf{x}}$, the **y** accuracy $\text{TSA}_{\mathbf{y}}$ for both the linear and Gaussian kernel. For simplicity, we denote the **x** accuracy in the linear setting as $\text{TSA}_{\mathbf{x}}(\text{L})$, while others are similarly defined.

The results are summarized in Figure 7. Four observations are worth highlighting. First, in both linear and Gaussian kernel settings, the smaller β_0 is, the higher the test accuracy for **x** becomes. This indicates that a bias can indeed be embedded in the classification boundary for the important class **x** by specifying a relatively smaller β_0. In comparison, MPM forces an equal treatment on each class and thus is not suitable for biased classification. Second, the test accuracies for **y** and **x** are strictly lower bounded by β_0 and α. This shows how a bias can be quantitatively, directly, and rigorously imposed towards the important class **x**. Note that again, for other weight-adapting based biased classifiers, the weights themselves lack accurate interpretations and thus cannot rigorously impose a specified bias, i.e., they would try different weights for a specified bias. Third, when given a prescribed β_0, the test accuracy for **x** and its worst-case accuracy α in the Gaussian kernel setting are both greater than the corresponding accuracies in the linear setting. Once again, this demonstrates the power of the kernelization. Fourth, we note that β_0 actually contains an upper bound, which is around 90% for the linear BMPM in this data set. This is reasonable. Observed from (7), the maximum β_0, denoted as β_{0_m}, is decided by setting $\alpha = 0$, i.e.,

$$\kappa(\beta_{0_m}) = \max_{\mathbf{a}\neq\mathbf{0}} \quad \frac{1}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}} \quad \text{s.t.} \quad \mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}}) = 1.$$

It is interesting to note that when β_0 is set to zero, the test accuracies for **y** in the linear and Gaussian settings are both around 50% (see Figure 7(b)). This seeming "irrationality" is actually reasonable. We will discuss this in the next section.

6. How Tight Is the Bound?

A natural question for MEMPM is, how tight is the worst-case bound? In this section, we present a theoretical analysis in addressing this problem.

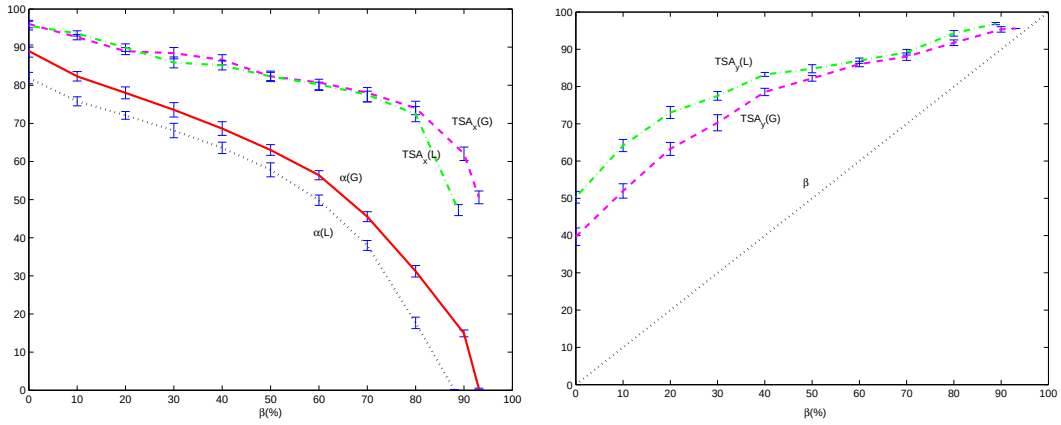
We begin with a lemma proposed in Popescu and Bertsimas (2001).

$$\sup_{\mathbf{y}\sim(\overline{\mathbf{y}},\Sigma_{\mathbf{y}})} \mathbf{Pr}\{\mathbf{y}\in\mathcal{S}\} = \frac{1}{1+d^2}, \quad \text{with} \quad d^2 = \inf_{\mathbf{y}\in\mathcal{S}} (\mathbf{y}-\overline{\mathbf{y}})^T\Sigma_{\mathbf{y}}^{-1}(\mathbf{y}-\overline{\mathbf{y}}), \tag{24}$$

where $\mathcal{S}$ denotes a convex set.

If we define $\mathcal{S} = \{\mathbf{a}^T\mathbf{y} \geq b\}$, the above lemma is changed to:

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



(a) (b)

Figure 7: Bounds and real accuracies. With β_0 varying from 0 to 1, the real accuracies are lower bounded by the worst-case accuracies. In addition, α(G) is above α(L), which shows the power of the kernelization.

By reference to (3), for a given hyperplane $\{\mathbf{a}, b\}$, we can easily obtain that

$$\beta = \frac{d^2}{1+d^2} \,. \tag{25}$$

Moreover, in Lanckriet et al. (2002b), a simple closed-form expression for the minimum distance d is derived:

$$d^2 = \inf_{\mathbf{a}^T\mathbf{y} \geq b} (\mathbf{y}-\overline{\mathbf{y}})^T \Sigma_{\mathbf{y}}^{-1} (\mathbf{y}-\overline{\mathbf{y}}) = \frac{\max((b-\mathbf{a}^T\overline{\mathbf{y}}),0)^2}{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \,. \tag{26}$$

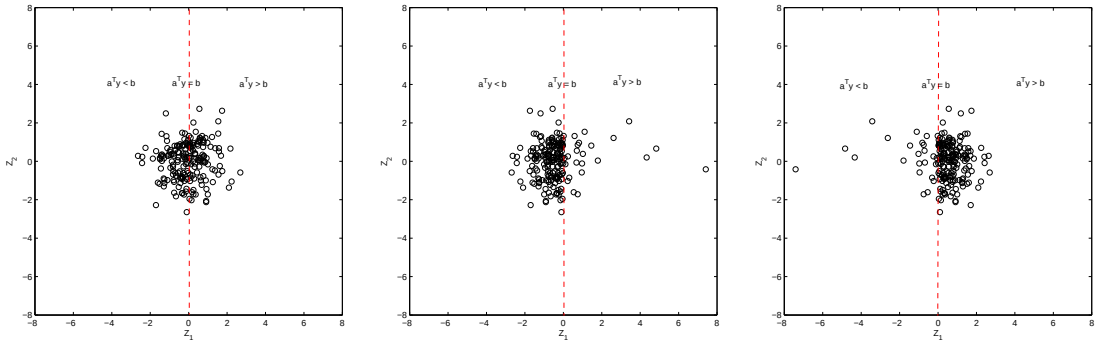
It is easy to see that when the decision hyperplane $\{\mathbf{a}, b\}$ passes the center $\overline{\mathbf{y}}$, d would be equal to 0 and the worst-case accuracy β would be 0 according to (25).

However, if we consider the Gaussian data (which we assume as $\mathbf{y}$ data) in Figure 8(a), a vertical line approximating $\overline{\mathbf{y}}$ would achieve about 50% test accuracy. The large gap between the worst-case accuracy and the real test accuracy seems strange. In the following, we construct an example of one-dimensional data to show the inner rationality of this observation. We attempt to provide the worst-case distribution containing the given mean and covariance, while a hyperplane passing its mean achieves a real test accuracy of zero.

Consider one-dimensional data y consisting of $N-1$ observations with values as m and one single observation with the value as $\sigma\sqrt{N}+m$. If we calculate the mean and the covariance, we obtain:

$$\overline{y} = m + \frac{\sigma}{\sqrt{N}} \,,$$
$$\Sigma_y = \frac{N-1}{N}\sigma^2 \,.$$

When N goes to infinity, the above one-dimensional data have the mean as m and the covariance as σ. In this extreme case, a hyperplane passing the mean will achieve a zero test accuracy, which is exactly the worst-case accuracy given the fixed mean and covariance as m and σ respectively. This example demonstrates the inner rationality of the minimax probability machines.



(a) (b) (c)

Figure 8: Three two-dimensional data sets with the same means and covariances but with different skewness. The worst-case accuracy bound of (a) is tighter than that of (b) and looser than that of (c).

To further examine the tightness of the worst-case bound in Figure 8(a), we vary β from 0 to 1 and plot against β the real test accuracy that a vertical line classifies the $\mathbf{y}$ data by using (25). Note that the real accuracy can be calculated as $\Phi(z \leq d)$. This curve is plotted in Figure 9.

Observed from Figure 9, the smaller the worst-case accuracy is, the looser it is. On the other hand, if we skew the $\mathbf{y}$ data towards the left side, while simultaneously maintaining the mean and covariance unchanged (see Figure 8(b)), an even bigger gap will be generated when β is small; similarly, if we skew the data towards the right side (see Figure 8(c)), a tighter accuracy bound will be expected. This finding means that adopting up to the second order moments only may not achieve a satisfactory bound. In other words, for a tighter bound, higher order moments such as skewness may need to be considered. This problem of estimating a probability bound based on moments is presented as the (n,k,Ω)-bound problem, which means "finding the tightest bound for an n-dimensional variable in the set Ω based on up to the k-th moments." Unfortunately, as proved in Popescu and Bertsimas (2001), it is NP-hard for $(n,k,\mathbb{R}^n)$-bound problems with $k \geq 3$. Thus tightening the bound by simply scaling up the moment order may be intractable in this sense. We may have to exploit other statistical techniques to achieve this goal. This certainly deserves a closer examination in the future.

7. On the Concavity of MEMPM

We address the issue of the concavity on the MEMPM model in this section. We will demonstrate that, although MEMPM cannot generally guarantee its concavity, there is strong empirical evidence showing that many real world problems demonstrate reasonable concavity in MEMPM. Hence, the

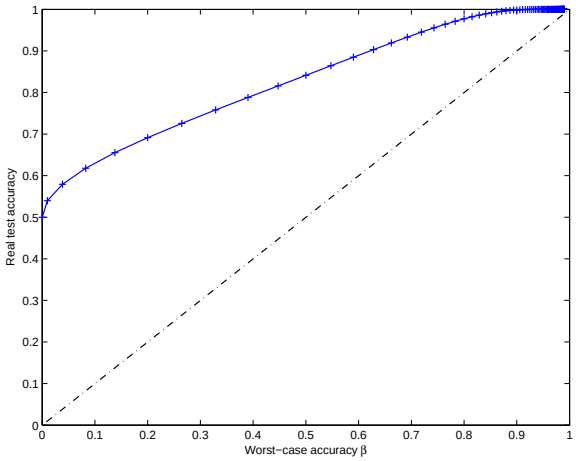


Figure 9: Theoretical comparison between the worst-case accuracy and the real test accuracy for the Gaussian data in Figure 8(a).

MEMPM model can be solved successfully by standard optimization methods, e.g., the linear search method proposed in this paper.

We first present a lemma for the BMPM model.

Lemma 8 *The optimal solution for BMPM is a strictly and monotonically decreasing function with respect to* β_0.

Proof Let the corresponding optimal worst-case accuracies on $\mathbf{x}$ be α_1 and α_2 respectively, when β_{01} and β_{02} are set to the acceptable accuracy levels for $\mathbf{y}$ in BMPM. We will prove that if $\beta_{01} > \beta_{02}$, then $\alpha_1 < \alpha_2$.

This can be proved by considering the contrary case, i.e., we assume $\alpha_1 \geq \alpha_2$. From the problem definition of BMPM, we have:

$$\begin{aligned} &\alpha_1 \geq \alpha_2 \Longrightarrow \kappa(\alpha_1) \geq \kappa(\alpha_2) \\ &\Longrightarrow \frac{1-\kappa(\beta_{01})\sqrt{\mathbf{a_1}^T\Sigma_\mathbf{y}\mathbf{a_1}}}{\sqrt{\mathbf{a_1}^T\Sigma_\mathbf{x}\mathbf{a_1}}} \geq \frac{1-\kappa(\beta_{02})\sqrt{\mathbf{a_2}^T\Sigma_\mathbf{y}\mathbf{a_2}}}{\sqrt{\mathbf{a_2}^T\Sigma_\mathbf{x}\mathbf{a_2}}}\,, \end{aligned} \tag{27}$$

where $\mathbf{a}_1$ and $\mathbf{a}_2$ are the corresponding optimal solutions that maximize $\kappa(\alpha_1)$ and $\kappa(\alpha_2)$ respectively, when β_{01} and β_{02} are specified.

From $\beta_{01} > \beta_{02}$ and (27), we have

$$\frac{1-\kappa(\beta_{02})\sqrt{\mathbf{a_1}^T\Sigma_\mathbf{y}\mathbf{a_1}}}{\sqrt{\mathbf{a_1}^T\Sigma_\mathbf{x}\mathbf{a_1}}} > \frac{1-\kappa(\beta_{01})\sqrt{\mathbf{a_1}^T\Sigma_\mathbf{y}\mathbf{a_1}}}{\sqrt{\mathbf{a_1}^T\Sigma_\mathbf{x}\mathbf{a_1}}} \geq \frac{1-\kappa(\beta_{02})\sqrt{\mathbf{a_2}^T\Sigma_\mathbf{y}\mathbf{a_2}}}{\sqrt{\mathbf{a_2}^T\Sigma_\mathbf{x}\mathbf{a_2}}}\,. \tag{28}$$

On the other hand, since $\mathbf{a}_2$ is the optimal solution of $\max_\mathbf{a} \frac{1-\kappa(\beta_{02})\sqrt{\mathbf{a}^T\Sigma_\mathbf{y}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_\mathbf{x}\mathbf{a}}}$, we have:

$$\frac{1-\kappa(\beta_{02})\sqrt{\mathbf{a_2}^T\Sigma_\mathbf{y}\mathbf{a_2}}}{\sqrt{\mathbf{a_2}^T\Sigma_\mathbf{x}\mathbf{a_2}}} \geq \frac{1-\kappa(\beta_{02})\sqrt{\mathbf{a_1}^T\Sigma_\mathbf{y}\mathbf{a_1}}}{\sqrt{\mathbf{a_1}^T\Sigma_\mathbf{x}\mathbf{a_1}}}\,.$$

This is obviously contradictory to (28). ■

From the sequential solving method of MEMPM, we know that MEMPM actually corresponds to a one-dimensional line search problem. More specifically, it further corresponds to maximizing the sum of two functions, namely, $f_1(\beta) + f_2(\beta)$,[4] where $f_1(\beta)$ is determined by the BMPM optimization and $f_2(\beta) = \beta$. According to Lemma 8, $f_1(\beta)$ strictly decreases as β increases. Thus it is strictly pseudo-concave. However, generally speaking, the sum of a pseudo-concave function and a linear function is not necessarily a pseudo-concave function and thus we cannot assure that every local optimum is the global optimum. This can be clearly observed in Figure 10. In this figure, f_1 is pseudo-concave in all three sub-figures; however, the sum $f_1 + f_2$ does not necessarily lead to a pseudo-concave function.

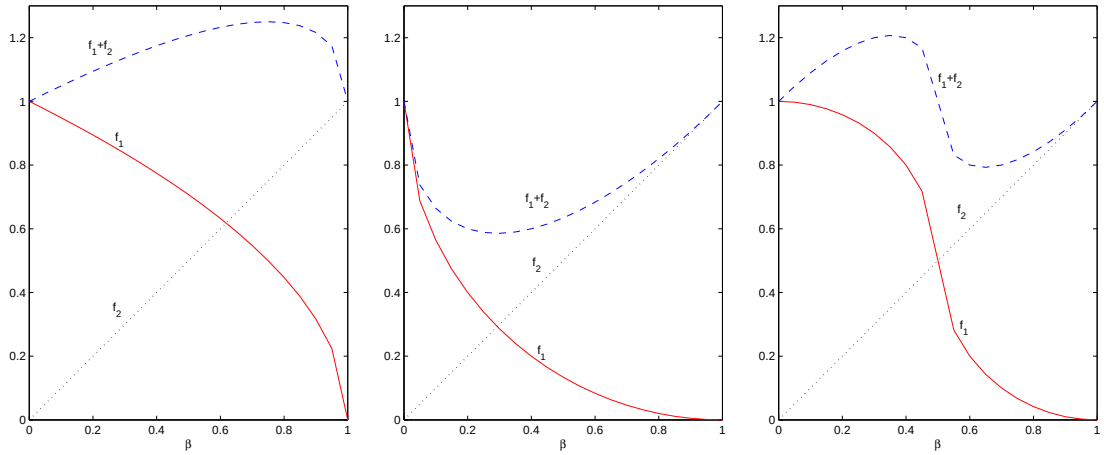
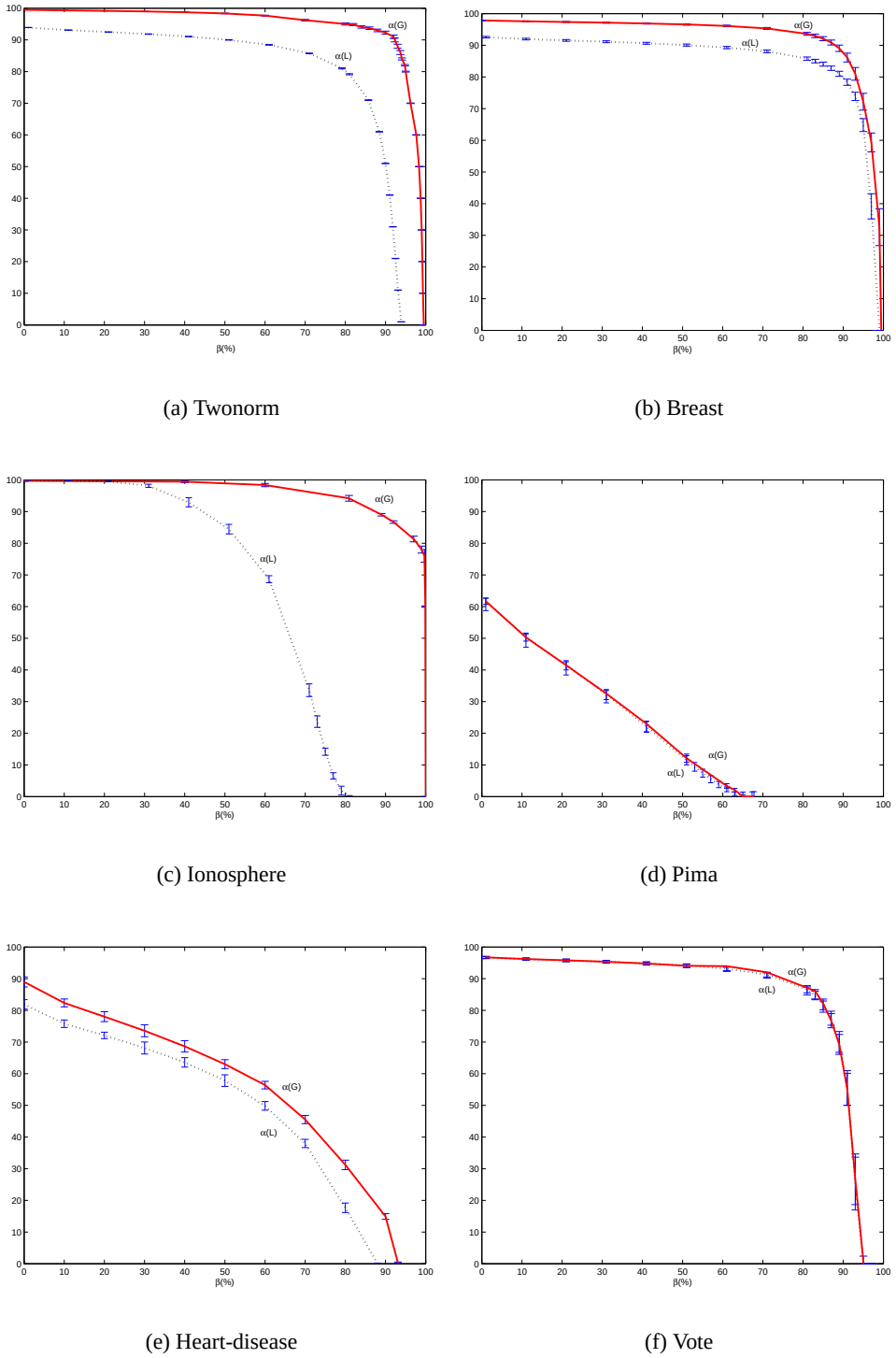


Figure 10: The sum of a pseudo-concave function and a linear function is not necessarily a concave function. In (a), $f_1 + f_2$ is a concave function, however in (b) and (c) it is not.

Nevertheless, there is strong empirical evidence showing that for many "well-behaved" real world classification problems, f_1 is overall concave, which results in the concavity of $f_1 + f_2$. This is first verified by the data sets used in this paper. We shift β from 0 to the corresponding upper bound and plot α against β in Figure 11. It is clearly observed that in all six data sets including both kernel and linear cases, the curves of α against β are overall concave. This motivates us to look further into the concavity of MEMPM. As shown in the following, when two classes of data are "well-separated," f_1 would be concave in the main "interest" region.

We analyze the concavity of $f_1(\beta)$ by imagining that β changes from 0 to 1. In this process, the decision hyperplane moves slowly from $\overline{\mathbf{y}}$ to $\overline{\mathbf{x}}$ according to (25) and (26). At the same time, $\alpha = f_1(\beta)$ should decrease accordingly. More precisely, if we denote $d_{\mathbf{x}}$ and $d_{\mathbf{y}}$ respectively as the

4. For simplicity, we assume θ as 0.5. Since a constant does not influence the concavity analysis, the factor of 0.5 is simply dropped.



(a) Twonorm (b) Breast

(c) Ionosphere (d) Pima

(e) Heart-disease (f) Vote

Figure 11: The curves of α against β (f_1) all tend to be concave in the data sets used in this paper.

Mahalanobis distances that $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ are from the associated decision hyperplane with a specified β, we can formulate the changing of α and β as

$$\alpha \to \alpha - k_1(d_{\mathbf{x}})\Delta d_{\mathbf{x}},$$
$$\beta \to \beta + k_2(d_{\mathbf{y}})\Delta d_{\mathbf{y}},$$

where $k_1(d_{\mathbf{x}})$ and $k_2(d_{\mathbf{y}})$ can be considered as the changing rate of α and β when the hyperplane lies $d_{\mathbf{x}}$ distance far away from $\overline{\mathbf{x}}$ and $d_{\mathbf{y}}$ distance far away from $\overline{\mathbf{y}}$ respectively. We consider the changing of α against β, namely, f_1':

$$f_1' = \frac{-k_1(d_{\mathbf{x}})\Delta d_{\mathbf{x}}}{k_2(d_{\mathbf{y}})\Delta d_{\mathbf{y}}}.$$

If we consider $d_{\mathbf{x}}$ and $\Delta d_{\mathbf{y}}$ insensitively change against each other or change with a proportional rate, i.e., $\Delta d_{\mathbf{x}} \approx c\Delta d_{\mathbf{y}}$ (c is a positive constant) as the decision hyperplane moves, the above equation can further be written as $f_1' = c\frac{-k_1(d_{\mathbf{x}})}{k_2(d_{\mathbf{y}})}$.

Lemma 9 *(1) If $d_{\mathbf{y}} \geq 1/\sqrt{3}$ or the corresponding $\beta \geq 0.25$, $k_2(d_{\mathbf{y}})$ decreases as $d_{\mathbf{y}}$ increases.*
(2) If $d_{\mathbf{x}} \geq 1/\sqrt{3}$ or the corresponding $\alpha \geq 0.25$, $k_1(d_{\mathbf{x}})$ decreases as $d_{\mathbf{x}}$ increases.

Proof Since (1) and (2) are very similar statements, we only prove (1). Note that $k_2(d)$ is the first order derivative of $\frac{d^2}{1+d^2}$ according to (25). We consider the first order derivative of $k_2(d)$ or the second order derivative of $\frac{d^2}{1+d^2}$. It is easily verified that $(\frac{d^2}{1+d^2})'' \leq 0$ when $d \geq 1/\sqrt{3}$. This is also illustrated in Figure 12. According to the definition of the second order derivative, we immediately obtain the lemma. Note that $d \geq 1/\sqrt{3}$ corresponds to $\beta \geq 0.25$. Thus the condition can also be replaced by $\beta \geq 0.25$.

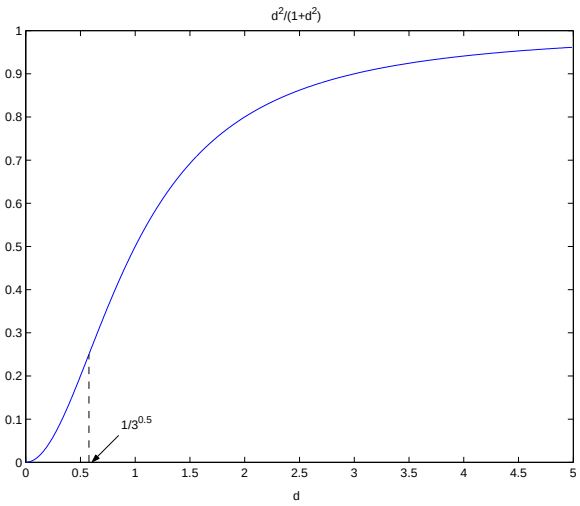


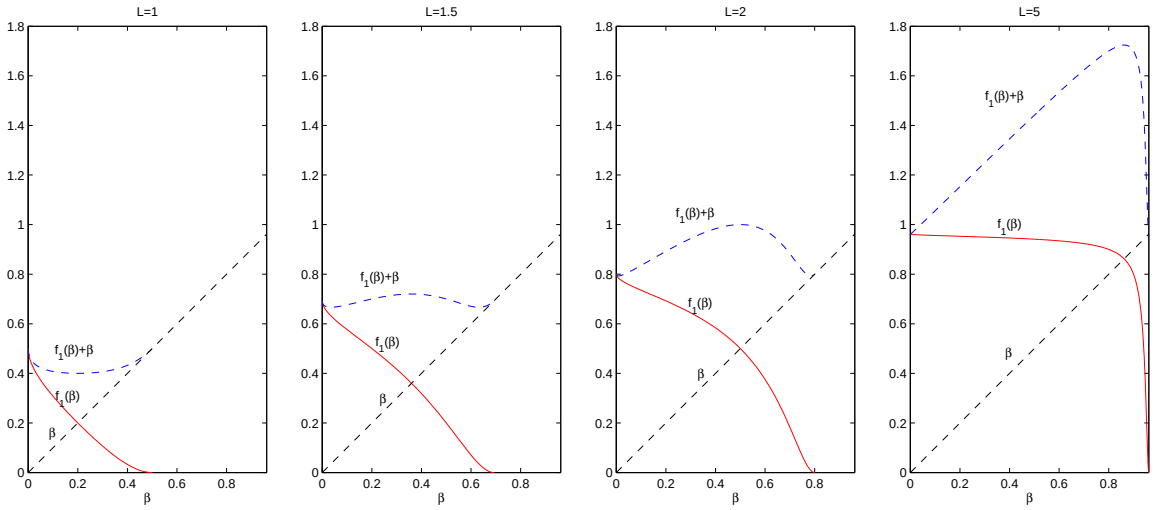
Figure 12: The curve of $d^2/(1+d^2)$. This function is concave when $d \geq 1/\sqrt{3}$.

■

In the above procedure, $d_{\mathbf{y}}$, β increase and $d_{\mathbf{x}}$, α decrease, as the hyperplane moves towards $\overline{\mathbf{x}}$.

Therefore, according to Lemma 9, $k_1(d_{\mathbf{x}})$ increases while $k_2(d_{\mathbf{y}})$ decreases when $\alpha, \beta \in [0.25,\ 1)$. This shows that f_1' is getting smaller as the hyperplane moves towards $\overline{\mathbf{x}}$. In other words, f_1'' would be less than 0, and it is concave when $\alpha, \beta \in [0.25,\ 1)$. It should be noted that in many well-separated real world data sets, there is a high possibility that the optimal α and β will be greater than 0.25, since to achieve good performance, the worst-case accuracies are naturally required to be greater than a certain small amount, e.g., 0.25. This is observed in the data sets used in the paper. All the data sets except the Pima data attain their optima satisfying this constraint. For Pima, the overall accuracy is relatively lower, which implies two classes of data in this data set appear to overlap substantially with each other.[5]

An illustration can also be seen in Figure 13. We generate two classes of Gaussian data with $\overline{\mathbf{x}} = [0,\ 0]^T$, $\overline{\mathbf{y}} = [L,\ 0]^T$, and $\Sigma_{\mathbf{x}} = \Sigma_{\mathbf{y}} = [1,\ 0; 0,\ 1]$. The prior probability for each data class is set to an equal value 0.5. We plot the curves of $f_1(\beta)$ and $f_1(\beta) + \beta$ when L is set to different values. It is observed that when two classes of data substantially overlap with each other, for example in Figure 13(a) with $L = 1$, the optimal solution of MEMPM lies in the small-value range of α and β, which is usually not concave. On the other hand, (b), (c), and (d) show that when two classes of data are well-separated, the optimal solutions lie in the region with $\alpha, \beta \in [0.25,\ 1)$, which is often concave.



(a) (b) (c) (d)

Figure 13: An illustration of the concavity of the MEMPM. Subfigure (a) shows that when two classes of data overlap substantially with each other, the optimal solution of MEMPM lies in the small-value range of α and β, which is usually not concave. (b), (c), and (d) show that when two classes of data are well-separated, the optimal solutions lie in the region with $\alpha, \beta \in [0.25,\ 1)$, which is often concave.

5. It is observed, even for Pima, the proposed solving algorithm is still successful, since α is approximately linear as shown in Figure 11. Moreover, due to the fact that the slope of α is slightly greater than -1, the final optimum naturally leads β to achieve its maximum.

Note that, in the above, we make an assumption that as the decision hyperplane moves, $d_{\mathbf{x}}$ and $d_{\mathbf{y}}$ change at an approximately fixed proportional rate. From the definition of $d_{\mathbf{x}}$ and $d_{\mathbf{y}}$, this assumption implies that $\mathbf{a}$, the direction of the optimal decision hyperplane, is insensitive to β. This assumption does not hold in all cases; however, observed from the geometrical interpretation of MEMPM, for those data with isotropic or not significantly anisotropic $\Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{y}}$, $\mathbf{a}$ would indeed be insensitive to β.

We summarize the above analysis in the following proposition.

Proposition 10 *Assuming (1) two classes of data are well-separated and (2) $d_{\mathbf{x}}$ and $d_{\mathbf{y}}$ change at an approximately fixed proportional rate as the optimal decision hyperplane (associated with a specified β) moves, the one-dimensional line search problem of MEMPM is often concave in the range of $\alpha, \beta \in [0.25, 1)$ and will often attain its optimum in this range. Therefore the proposed solving method leads to a satisfactory solution.*

Remarks. As demonstrated in the above, although the MEMPM is often overall concave in real world tasks, there exist cases that the MEMPM optimization problem is not concave. This may lead to a local optimum, which may not be the global optimum. In this case, we may need to choose the initial starting point carefully. In addition, the physical interpretation of β as the worst-case accuracy may make it relatively easy to choose a suitable initial value. For example, we can set the initial value by using the information obtained from prior domain knowledge.

8. Limitations and Future Work

In this section, we present the limitations and future work. First, although MEMPM achieves better performance than the MPM, its sequential optimization of the Biased Minimax Probability Machine may cost more training time than MPM. Although in pattern recognition tasks, especially in off-line classification, effectiveness is often more important than efficiency, expensive time-cost presents one of the main limitations of the MEMPM model, in particular for large scale data sets with millions of samples. To solve this problem, one possible direction is to eliminate those redundant points that make less contribution to the classification. In this way, the problem dimension (in the kernelization) would be greatly decreased and this may help in reducing the computational time required. Another possible direction is to exploit some techniques to decompose the Gram matrix (as is done in SVM) and to develop some specialized optimization procedures for MEMPM. Recently, we also note that Strohmann et al. (2004) have proposed a speed-up method by exploiting the sparsity of MPM. Undoubtedly, speeding up the algorithm will be a highly worthy topic in the future.

Second, as a generalized model, MEMPM actually incorporates some other variations. For example, when the prior probability (θ) cannot be estimated reliably (e.g., in sparse data), maximizing $\alpha + \beta$, namely the sum of the accuracies or the difference between true positive and false positive, would be considered. This scheme is widely used in the pattern recognition field, e.g., in medical diagnosis (Grzymala-Busse et al., 2003) and in graph detection, especially line detection and arc detection, where it is called the Vector Recovery Index (Liu and Dori, 1997; Dori and Liu, 1999). Moreover, when there are domain experts at hand, a variation of MEMPM, namely, the maximization of $C_{\mathbf{x}}\alpha + C_{\mathbf{y}}\beta$ may be used, where $C_{\mathbf{x}}$ ($C_{\mathbf{y}}$) is the cost of a misclassification of $\mathbf{x}$ ($\mathbf{y}$) obtained from experts. Exploring these variations in some specific domains is thus a valuable direction in the future.

Third, we have proposed a general framework for robustly estimating model input parameters, namely, the means and covariances. Based on this framework, estimating the input vector or matrix parameters is changed to finding four adapting scale parameters, i.e., $\nu_{\mathbf{x}}, \nu_{\mathbf{y}}, \rho_{\mathbf{x}}$, and $\rho_{\mathbf{y}}$. While we may obtain these four parameters by conducting cross validation in small data sets, it is computationally hard to do this in large scale data sets. Although one possible way to determine these values is based on the central limit theorem or the resampling method (Lanckriet et al., 2002b), it is still valuable to investigate other techniques in the future.

Fourth, Lanckriet et al. (2002b) have built up a connection between MPM and SVM from the perspective of the margin definition, i.e., MPM corresponds to finding the hyperplane with the maximal margin from the class center. Nevertheless, some deeper connections need to be investigated, e.g., how is the bound of MEMPM related to the generalization bound of SVM? More recently, Huang et al. (2004a) have disclosed the relationship between them from either a local or a global viewpoint of data. It is particularly useful to look into these links and explore their further connections in the future.

9. Conclusion

The Minimax Probability Machine achieves performance in classification tasks that is comparable to that of a state-of-the-art classifier, the Support Vector Machine. This model attempts to minimize the worst-case probability of misclassification of future data points. However, its equality constraint on the worst-case accuracies for two classes makes it unnecessarily minimize the error rate in the worst-case setting and thus cannot assure the optimal classifier in this sense.

In this paper, we have proposed a generalized Minimax Probability Machine, called the Minimum Error Minimax Probability Machine, which removes the equality constraint on the worst-case accuracies for two classes. By minimizing the upper bound of the Bayes error of future data points, our approach derives the distribution-free Bayes optimal hyperplane in the worst-case setting. More importantly, we have shown that the worst-case Bayes optimal hyperplane derived by MEMPM becomes the true Bayes optimal hyperplane when certain conditions are satisfied, in particular, when Gaussianity is assumed for the data. We have evaluated our algorithm on both synthetic data sets and real world benchmark data sets. The performance of MEMPM is demonstrated to be very promising. Moreover, the validity of our proposition, i.e., the minimum error rate Minimax Probability Machine is not certain to achieve the same worst-case accuracies for two classes, has also been verified by the experimental results.

Acknowledgements

We thank Gert R. G. Lanckriet at U.C. Berkeley for providing the Matlab source code of Minimax Probability Machine on the web. We extend our thanks to the editor and the anonymous reviewers for their valuable comments. The work described in this paper was fully supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4182/03E and Project No. CUHK4235/04E).

References

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, Massachusetts, 2nd edition, 1999.

C. L. Blake and C. J. Merz. Repository of machine learning databases, University of California, Irvine, http://www.ics.uci.edu/∼mlearn/MLRepository.html, 1998.

L. Breiman. Arcing classifiers. Technical Report 460, Statistics Department, University of California, 1997.

Y. S. Chow and H. Teicher. *Probability theory: Independence, interchageability, martingales*. Springer-Verlag, New York, 3rd edition, 1997.

B. D. Craven. *Mathematical Programming and Control Theory*. Chapman and Hall, London, 1978.

B. D. Craven. *Fractional Programming, Sigma Series in Applied Mathematics 4*. Heldermann Verlag, Berlin, 1988.

G. Deco and D. Obradovic. *An information-theoretic approach to neural computing*. Springer-Verlag, 1996.

D. Dori and W. Liu. Sparse pixel vectorization: An algorithm and its performance evaluation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21:202–215, 1999.

J. W. Grzymala-Busse, L. K. Goodwin, and X. Zhang. Increasing sensitivity of preterm birth by changing rule strengths. *Pattern Recognition Letters*, 24:903–910, 2003.

K. Huang, H. Yang, I. King, and M. R. Lyu. Learning large margin classifiers locally and globally. In *The Twenty-First International Conference on Machine Learning (ICML-2004)*, pages 401–408, 2004a.

K. Huang, H. Yang, I. King, M. R. Lyu, and L. Chan. Biased minimax probability machine for medical diagnosis. In *the Eighth International Symposium on Artificial Intelligence and Mathematics (AMAI-2004)*, 2004b.

D. Keysers, F. J. Och, and H. Ney. Maximum entropy and gaussian models for image object recognition. In *Proceedings of the Twenty-Fourth DAGM Symposium, Volume 2449 of Lecture Notes in Computer Science*, pages 498–506. Springer-Verlag, 2002.

G. R. G. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan. Minimax probability machine. In *Advances in Neural Information Processing Systems (NIPS 14)*, 2002a.

G. R. G. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3:555–582, 2002b.

W. Liu and D. Dori. A protocol for performance evaluation of line detection algorithms. *Machine Vision and Application*, 9:240–250, 1997.

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

M. A. Maloof, P. Langley, T. O. Binford, R. Nevatia, and S. Sage. Improved rooftop detection in aerial images with machine learning. *Machine Learning*, 53:157–191, 2003.

Y. Nesterov and A. Nemirovsky. *Interior point polynomial methods in convex programming: Theory and applications*. Studies in Applied Mathematics 13. Society for Industrial & Applied Math, 1994.

E. Osuna, R. Freund, and F. Girosi. Support Vector Machines: Training and Applications. Technical Report AIM-1602, MIT, 1997. URL `citeseer.nj.nec.com/osuna97support.html`.

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

S. Schaible. Fractional programming. *Zeitschrift für Operational Research*, Serie A 27(1):39–54, 1977.

S. Schaible. Fractional programming. In R. Horst and P. M. Pardalos, editors, *Handbook of Global Optimization*, Nonconvex Optimization and its Applications, pages 495–608. Kluwer Academic Publishers, Dordrecht-Boston-London, 1995.

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

T. Strohmann, A. Belitski, G. Grudic, and D. DeCoste. Sparse greedy minimax probability machine classification. In *Advances in Neural Information Processing Systems (NIPS 16)*. 2004.

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