Second-Order Measures of Quality for Binary Classification: a Critical Overview and Their Use for Nonlinear Receiver Design

Fahed Abdallah
Cédric Richard
Régis Lengellé
Université de Technologie de Troyes (UTT), France

When deriving a detector, we are often led to consider design criteria such as second-order measures of quality. The aim of this paper is to provide a critical overview of these criteria. We first consider the case of deriving unconstrained detectors. We show that second-order criteria must satisfy a non-trivial condition to yield Bayes-optimal receivers, to be considered as relevant criteria for detector design. Next, we address the case where constraints are imposed on the detection structure, leading us to consider some set $\mathcal{D}$ of admissible detectors. In these conditions we prove that even if there exists a monotonic function of the likelihood ratio in $\mathcal{D}$, obtaining this statistic via the optimization of a second-order criterion, relevant or not, is not guaranteed. Results are illustrated by simulation examples. Finally, in order to derive nonlinear discriminants via optimization of second-order criteria, we propose a method based on the kernel trick used in the implementation of the well-known support vector machine method. The new method is tested on a number of real data sets.

Keywords: Maximum likelihood, Detection, Distance measures , Signal-to-noise ratio, nonlinear discriminants.

INTRODUCTION

Let $(\boldsymbol{X}, Y)$ be a pair of random variables taking taking their respective values from $\mathbb{R}^d$ and $\{0, 1\}$, where $\boldsymbol{X}$ is the observation and Y indicate either class $\mathcal{C}_0$ or $\mathcal{C}_1$. The purpose of detection is to determine to which of two classes ($Y = 0$ or $Y = 1$) a given observation $\boldsymbol{X}$ belongs. According to classical statistical detection theories, comparing any strictly monotonic function of the likelihood ratio $L(\boldsymbol{X})$ with a threshold value is the optimum test [13]. In practical applications, implementing such a test may be impossible because of incomplete specification of the conditional probability densities $p(\boldsymbol{X}|Y = 0)$ and $p(\boldsymbol{X}|Y = 1)$, denoted by the standard notations $p_0(\boldsymbol{X})$ and $p_1(\boldsymbol{X})$, respectively. Therefore we are often led to consider alternative design criteria such as second-order measures of quality. These criteria are easy to use since they only depend on first and second-order moments of the statistics S to be sought [6, 7]. A wide variety of second-order measures of performance have been proposed and several contributions have been presented to prove their efficiency, e.g., [8] and references therein.

In particular, some of these criteria guarantee the best solution in the Bayes sense since there optimization leads to a monotonic function of the likelihood ratio, as has been shown for well-known criteria such as Fisher criterion, mean-square error and signal-to-noise ratio.

The aim of this paper is to provide an overview of the strengths and shortcomings of second-order criteria. First, we consider the case of deriving unconstrained detectors. We show that second-order criteria must satisfy a non-trivial condition to provide Bayes-optimal receivers, to be considered as relevant second-order criteria for detector design. Next, we address the case where constraints are imposed on the structure of the detector, leading us to restrict our attention to some set $\mathcal{D}$ of admissible detectors. In these conditions, we prove that even if a monotonic function of the likelihood ratio in $\mathcal{D}$ exists, obtaining this statistic via the optimization of a relevant criterion is not guaranteed. Finally, an original method to derive nonlinear discriminants via optimization of second-order criteria is introduced. It uses the same kernel trick as the support vector machines (SVMs). It allows us to develop a nonlinear generalization of linear receivers obtained

via the optimization of second-order criteria. Simulation results are presented in order to compare the resulting nonlinear method with that of SVMs.

CHARACTERIZATION OF RELEVANT SECOND-ORDER CRITERIA

Background And Notations

Let $S(\boldsymbol{X}) : \mathbb{R}^d \longrightarrow \mathbb{R}$ be an arbitrary measurable function and let $g : \mathbb{R}^d \longrightarrow \{0, 1\}$ be the decision function based on $S(\boldsymbol{X})$:

$$g(\boldsymbol{X}) = \begin{cases} 1 & \text{if } S(\boldsymbol{X}) > b \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

which errs on $\boldsymbol{X}$ if $g(\boldsymbol{X}) \neq Y$. Classical statistical theories such as Bayes, Neyman-Pearson and minimax lead to the fundamental result that the optimum test consists of comparing the likelihood ratio $L(\boldsymbol{X}) \triangleq p_1(\boldsymbol{X})/p_0(\boldsymbol{X})$ with a given threshold b in order to make a decision [13]. The decision rule can then be expressed as

$$g^*(\boldsymbol{X}) = \begin{cases} 1 & \text{if } L(\boldsymbol{X}) > b \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Note that g is equivalent[1] to the Bayes-optimal detector g^*, which means that their receiver operating characteristic (ROC) is the same, if $S(\boldsymbol{X}) = \phi\{L(\boldsymbol{X})\}$, where ϕ is any monotonic function. Since the implementation of (2) may be impossible in many practical applications, we are often led to consider simpler procedures for designing (1). In particular, one can use alternative design criteria such as second-order measures of performance. These criteria are defined in terms of first and second-order moments of the statistics $S(\boldsymbol{X})$, namely

$$m_i \triangleq \mathrm{E}\{S \,|\, Y = i\}, \qquad \sigma_i^2 \triangleq \mathrm{Var}\{S \,|\, Y = i\}, \tag{3}$$

with $i \in \{0, 1\}$. There have been many contributions to justify the use of individual second-order criteria, including convergence with optimal detectors of classical detection theories (see, e.g., [8] and references therein). Examples of these criteria are Fisher, mean-square error and signal-to-noise ratio. In [7, pp. 141-3], the objective of the author is to unify these results stating that the use of any function $\Psi(m_0, m_1, \sigma_0^2, \sigma_1^2)$ as a criterion for general non-linear detector design leads to a Bayes-optimum detector. In fact, we show in the next subsection that $\Psi(m_0, m_1, \sigma_0^2, \sigma_1^2)$ must satisfy a non-trivial condition to guarantee Bayes-optimal detectors for general nonlinear detector design, and thus to be considered as a *relevant second-order criterion* for detector design.

[1] Throughout this paper, two detectors are said to be equivalent if their receiver operating characteristic are the same.

Relevant Second-Order Criteria

Let Ψ be any function of $m_i \triangleq \int S(\boldsymbol{X})\, p_i(\boldsymbol{X})\, d\boldsymbol{X}$ and $\sigma_i^2 \triangleq \int (S(\boldsymbol{X}) - m_i)^2\, p_i(\boldsymbol{X})\, d\boldsymbol{X}$, where $S(\boldsymbol{X})$ denotes any decision statistic. We first have to characterize statistics $S(\boldsymbol{X})$ which optimize Ψ. Operating on Ψ with a variational calculus, we obtain

$$\delta\Psi = \frac{\partial\Psi}{\partial m_0}\,\delta m_0 + \frac{\partial\Psi}{\partial m_1}\,\delta m_1 + \frac{\partial\Psi}{\partial \sigma_0^2}\,\delta\sigma_0^2 + \frac{\partial\Psi}{\partial \sigma_1^2}\,\delta\sigma_1^2. \tag{4}$$

Since $\delta m_i = \int \delta S(\boldsymbol{X})\, p_i(\boldsymbol{X})\, d\boldsymbol{X}$ and $\delta\sigma_i^2 = 2\int (S(\boldsymbol{X}) - m_i)\,\delta S(\boldsymbol{X})\, p_i(\boldsymbol{X})\, d\boldsymbol{X}$ with $i \in \{0, 1\}$, we obtain

$$\begin{aligned} \delta\Psi = \int \Big[& \frac{\partial\Psi}{\partial m_0}\, p_0(\boldsymbol{X}) + \frac{\partial\Psi}{\partial m_1}\, p_1(\boldsymbol{X}) \\ & + 2\,(S(\boldsymbol{X}) - m_0)\,\frac{\partial\Psi}{\partial \sigma_0^2}\, p_0(\boldsymbol{X}) \\ & + 2\,(S(\boldsymbol{X}) - m_1)\,\frac{\partial\Psi}{\partial \sigma_1^2}\, p_1(\boldsymbol{X})\Big]\delta S(\boldsymbol{X})\, d\boldsymbol{X}. \end{aligned} \tag{5}$$

In order to make $\delta\Psi = 0$ regardless of $\delta S(\boldsymbol{X})$, the $[\cdot]$ term in the integrand must be equal to 0. Using $L(\boldsymbol{X}) = \frac{p_1(\boldsymbol{X})}{p_0(\boldsymbol{X})}$, we finally get the expression of the statistic $S(\boldsymbol{X})$ optimizing Ψ as a function of the likelihood ratio

$$S(\boldsymbol{X}) = -\frac{1}{2}\,\frac{\frac{\partial\Psi}{\partial m_0} + \frac{\partial\Psi}{\partial m_1} L(\boldsymbol{X})}{\frac{\partial\Psi}{\partial \sigma_0^2} + \frac{\partial\Psi}{\partial \sigma_1^2} L(\boldsymbol{X})} + \frac{m_0 \frac{\partial\Psi}{\partial \sigma_0^2} + m_1 \frac{\partial\Psi}{\partial \sigma_1^2} L(\boldsymbol{X})}{\frac{\partial\Psi}{\partial \sigma_0^2} + \frac{\partial\Psi}{\partial \sigma_1^2} L(\boldsymbol{X})} \tag{6}$$

The above statistic $S(\boldsymbol{X})$ leads to a Bayes-optimal detector if, and only if, it is a strictly monotonic function of $L(\boldsymbol{X})$. Evaluating the first order derivative of $S(\boldsymbol{X})$ with respect to $L(\boldsymbol{X})$, we obtain

$$\frac{dS}{dL}(\boldsymbol{X}) = \frac{(m_1 - m_0)\frac{\partial\Psi}{\partial \sigma_0^2}\frac{\partial\Psi}{\partial \sigma_1^2} + \frac{1}{2}\left(\frac{\partial\Psi}{\partial \sigma_1^2}\frac{\partial\Psi}{\partial m_0} - \frac{\partial\Psi}{\partial \sigma_0^2}\frac{\partial\Psi}{\partial m_1}\right)}{\left(\frac{\partial\Psi}{\partial \sigma_0^2} + \frac{\partial\Psi}{\partial \sigma_1^2} L(\boldsymbol{X})\right)^2} \tag{7}$$

We thus note that $S(\boldsymbol{X})$ defined by (6) is a strictly monotonic function of $L(\boldsymbol{X})$ if, and only if, the numerator of (7) is not equal to 0. This result leads directly to the following proposition [14].

Proposition 1. *$\Psi(m_0, m_1, \sigma_0^2, \sigma_1^2)$ is a relevant second-order criterion Ψ_R, i.e., it guarantees the best solution in the Bayes sense, if and only if*

$$(m_1 - m_0)\,\frac{\partial\Psi}{\partial \sigma_0^2}\,\frac{\partial\Psi}{\partial \sigma_1^2} + \frac{1}{2}\left(\frac{\partial\Psi}{\partial \sigma_1^2}\,\frac{\partial\Psi}{\partial m_0} - \frac{\partial\Psi}{\partial \sigma_0^2}\,\frac{\partial\Psi}{\partial m_1}\right) \neq 0. \tag{8}$$

Since it is very difficult, if not impossible, to find the solutions of (8), the above property can only be used to test the relevance of any criteria $\Psi(m_0, m_1, \sigma_0^2, \sigma_1^2)$. However, note that (8) is a non-restrictive condition, i.e., there exists a broad class of second-order criteria that lead to Bayes-optimal detectors [1].

CONSTRAINED DETECTOR DESIGN USING SECOND-ORDER CRITERIA

As shown the previous section, there exists a broad class of second-order criteria that lead to a detection statistic $S(\boldsymbol{X})$ equivalent to the likelihood ratio $L(\boldsymbol{X})$. However, implementing $S(\boldsymbol{X})$ remains an unsolved problem since it depends on the probability densities $p_0(\boldsymbol{X})$ and $p_1(\boldsymbol{X})$ via $L(\boldsymbol{X})$, which are unknown. Therefore, we have to use the following strategy for deriving receivers [5]:

1. selecting a class $\mathcal{D}$ of detection statistics;
2. choosing the statistic of $\mathcal{D}$ that optimizes a given measure of performance, e.g., a relevant second-order criterion.

Unfortunately, this approach does not necessarily provide a Bayes-optimal detector since it generally requires the optimum statistic (6) to be a member of $\mathcal{D}$. We shall now discuss this drawback in the case where $\mathcal{D}$ denotes the class of linear statistics.

Linear Detectors Design

Linear detectors are the simplest to use as far as their implementation is concerned, and are directly related to many known techniques such as correlations and Euclidean distances [7]. We shall now show how second-order criteria can be used for designing linear detectors, which are defined as follows:

$$g(\boldsymbol{X}) = \begin{cases} 1 & \text{if } S(\boldsymbol{X}) = \boldsymbol{W}^T\boldsymbol{X} - b > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Here $\boldsymbol{W}$ denotes the direction onto which any n-dimensional observation $\boldsymbol{X}$ is projected, and b is the detector threshold. The conditional expected values and variances of $S(\boldsymbol{X})$ are given by

$$m_i = \mathrm{E}\{S \,|\, Y = i\} = \boldsymbol{W}^T\mathbf{M}_i - b \tag{10}$$

$$\sigma_i^2 = \mathrm{Var}\{S \,|\, Y = i\} = \boldsymbol{W}^T\boldsymbol{\Sigma}_i\boldsymbol{W}, \tag{11}$$

where $\mathbf{M}_i$ and $\boldsymbol{\Sigma}_i$ are the conditional expected vectors and covariance matrices of $\boldsymbol{X}$. Let Ψ be any second-order criterion. The optimal statistic $S(\boldsymbol{X})$ is given by equating to zero the partial derivatives of Ψ with respect to $\boldsymbol{W}$ and b. As shown in [7, pp. 133-4] and [12], solving this linear system leads directly to the following proposition.

Proposition 2. *Let $S(\boldsymbol{X}) \triangleq \boldsymbol{W}^T\boldsymbol{X} - b$ be any linear decision statistic. The optimum projection vector $\boldsymbol{W}$ under which the maximum value of any second-order criteria Ψ is reached satisfies*

$$\boldsymbol{W}_\rho = [\rho\boldsymbol{\Sigma}_0 + (1-\rho)\boldsymbol{\Sigma}_1]^{-1}\,[\mathbf{M}_1 - \mathbf{M}_0], \tag{12}$$

where $\mathbf{M}_i$ and $\boldsymbol{\Sigma}_i$ are the conditional expected vectors and covariance matrices of $\boldsymbol{X}$. The parameter ρ depends on the criterion Ψ as follows:

$$\rho = \frac{\frac{\partial\Psi}{\partial\sigma_0^2}}{\frac{\partial\Psi}{\partial\sigma_0^2} + \frac{\partial\Psi}{\partial\sigma_1^2}}. \tag{13}$$

The optimum projection direction $\boldsymbol{W}_\rho$ depends on Ψ through a single parameter $\rho \in]-\infty, +\infty[$. In this case, the latter can be chosen to optimize the performance of the detector. Note that $\rho \in [0, 1]$ if, and only if, $\partial\Psi/\partial\sigma_0^2$ and $\partial\Psi/\partial\sigma_1^2$ are of the same sign (**Property 1**). This condition means that Ψ varies in the same way with σ_0^2 and σ_1^2, which is a desirable but non-mandatory requirement for design criteria. Let us now concentrate on $\boldsymbol{W}_{-\infty}$ and $\boldsymbol{W}_{+\infty}$. They are both proportional to $[\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1]^{-1}[\mathbf{M}_1 - \mathbf{M}_0]$ since we have

$$\boldsymbol{W}_{\pm\infty} \propto \lim_{\rho\to\pm\infty} \frac{1}{\rho}[\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1]^{-1}[\mathbf{M}_1 - \mathbf{M}_0]. \tag{14}$$

The projection directions $\boldsymbol{W}_{-\infty}$ and $\boldsymbol{W}_{+\infty}$ then lead to equivalent detection structures (**Property 2**) since the ROC depends only on the direction of $\boldsymbol{W}$.

In the following, Proposition 2 and the above properties are illustrated through some classic detection problems. As mentioned at the very beginning of this section, we also show that (relevant) second-order criteria does not guarantee an optimal detector in the Bayes sense if we restrict the solution space to a specific class $\mathcal{D}$ of detectors. Here this drawback is illustrated through the following situations:

Scenario 1: the optimization of any second-order criteria in $\mathcal{D}$ leads to a Bayes-optimal detector,

Scenario 2: certain second-order criteria exist which provide Bayes-optimal receivers in $\mathcal{D}$,

Scenario 3: there exists a Bayes-optimal detector in $\mathcal{D}$ but it cannot be reached by optimizing any second-order criteria,

where $\mathcal{D}$ denotes the class of linear detectors (9).

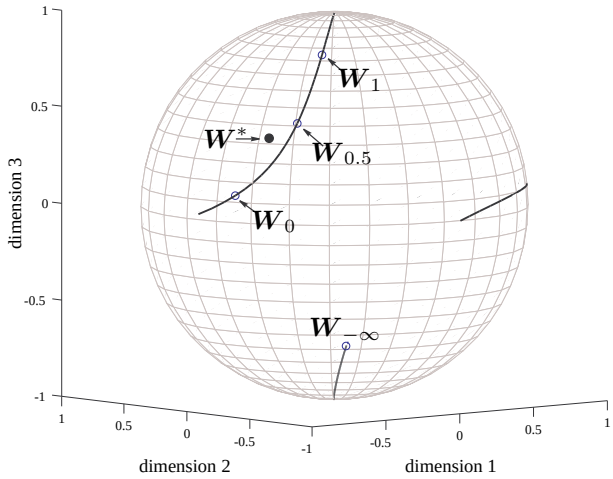

Fig. 1. Evolution of the projection direction $\boldsymbol{W}_\rho$ on the unit sphere as a function of $\rho \in]-\infty, +\infty[$, in the case of three-dimensional exponential distributions ($\lambda_{01} = 5$, $\lambda_{11} = 2$, $\lambda_{02} = 3$, $\lambda_{12} = 2$, $\lambda_{03} = 2$, $\lambda_{13} = 3$). The projection direction associated with the Bayes-optimal detector is referred to as $\boldsymbol{W}^*$ in this figure.

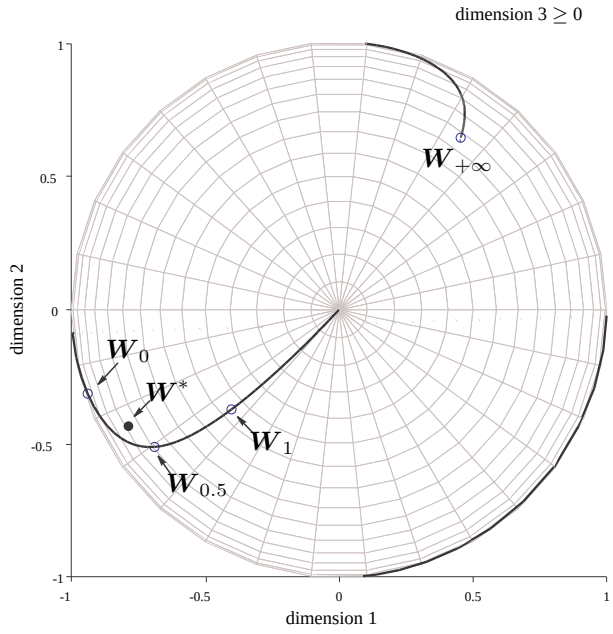

Fig. 2. Same as figure 1. Overhead view of the unit sphere, i.e., dimension $3 \geq 0$.

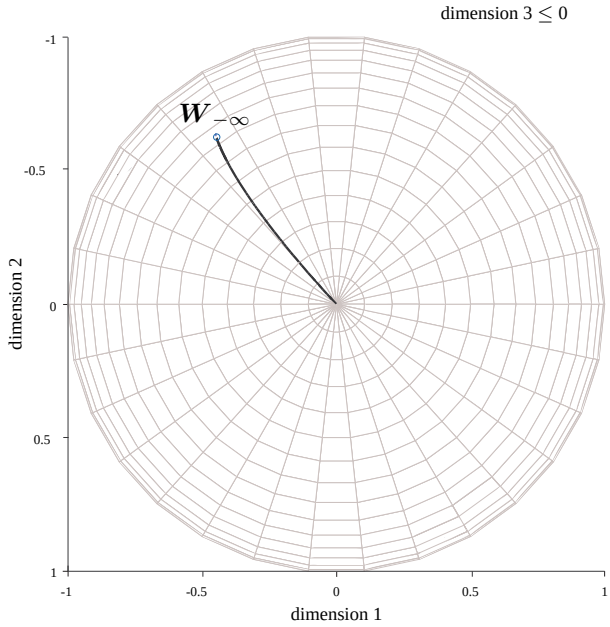

Fig. 3. Same as figure 1. View from below of the unit sphere, i.e., dimension $3 \leq 0$.

Data set	KFD	SVM	KSOD
Banana	10.60	**10.43** (132)	10.59
Thyroid	0.39	0.33 (156)	**0.25**
B. cancer	8.14	7.10 (364)	**6.70**
Diabetes	17.79	17.68 (308)	**17.39**
German	21.36	21.06 (400)	**20.96**
Heart	4.44	4.52 (388)	**4.41**
Solar	32.42	32.73 (364)	**31.61**
Waveform	11.14	**11.07** (400)	11.14
Ringnorm	1.53	**1.52** (396)	1.53
Titanic	28.88	28.88 (400)	28.55

Table 1. Comparison of the mean error rates obtained with KFD, SVM and KSOD. This table also gives the number of support vectors used by the SVM method.

Scenario 1: Case Of Normal Distributions With Equal Covariances

When $p_0(\boldsymbol{X})$ and $p_1(\boldsymbol{X})$ are normal with expected vectors $\mathbf{M}_0$ and $\mathbf{M}_1$ and covariance matrices $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$, it is well-known that the Bayes-optimal statistic is given by:

$$S(\boldsymbol{X}) = \frac{1}{2}(\boldsymbol{X}-\mathbf{M}_0)^T\mathbf{\Sigma}_0^{-1}(\boldsymbol{X}-\mathbf{M}_0) - \frac{1}{2}(\boldsymbol{X}-\mathbf{M}_1)^T\mathbf{\Sigma}_1^{-1}(\boldsymbol{X}-\mathbf{M}_1). \quad (15)$$

This equation shows that the decision boundary is a quadratic form in $\boldsymbol{X}$. When $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1 = \mathbf{\Sigma}$, the boundary becomes a linear function of $\boldsymbol{X}$ as

$$S(\boldsymbol{X}) = (\mathbf{M}_1-\mathbf{M}_0)^T\mathbf{\Sigma}^{-1}\boldsymbol{X}. \quad (16)$$

Eq. (16) indicates that the direction onto which any $\boldsymbol{X}$ is projected is given by $\boldsymbol{W}^* = \mathbf{\Sigma}^{-1}(\mathbf{M}_1-\mathbf{M}_0)$. Let us now determine the projection direction under which the optimal value of any second-order criterion is reached. Eq. (12) gives:

$$\boldsymbol{W}_\rho = \mathbf{\Sigma}^{-1}(\mathbf{M}_1-\mathbf{M}_0). \quad (17)$$

Comparing the projection directions $\boldsymbol{W}^*$ and $\boldsymbol{W}_\rho$, we immediately conclude that they both correspond to equivalent detection structures. This very simple example shows that any second-order criterion, relevant or not, can sometimes lead to a Bayes-optimal receiver. However, such a success is not always guaranteed as illustrated in the next subsection, even if it exists a Bayes-equivalent receiver in $\mathcal{D}$.

Scenarios 2 And 3: Case Of Exponential Distributions

Let us consider that the components X_j of $\boldsymbol{X}$ are exponentially distributed and mutually independent. Then we have:

$$p_i(\boldsymbol{X}) = \prod_{j=1}^{d}\frac{1}{\lambda_{ij}}\exp\left(-\frac{1}{\lambda_{ij}}X_j\right)u(X_j), \qquad i\in\{0,1\}, \quad (18)$$

with λ_{ij} the parameter of the exponential distribution of the random variable X_j, and $u(.)$ the step function. It can easily be shown that the linear function given below is the Bayes-optimal detection statistic:

$$S(\boldsymbol{X}) = \sum_{j=1}^{d}\left(\frac{1}{\lambda_{0j}}-\frac{1}{\lambda_{1j}}\right)X_j. \quad (19)$$

Then it is associated with the following Bayes-optimal projection direction:

$$\boldsymbol{W}^* = \left(\frac{\lambda_{11}-\lambda_{01}}{\lambda_{11}\lambda_{01}},\ldots,\frac{\lambda_{1j}-\lambda_{0j}}{\lambda_{1j}\lambda_{0j}},\ldots,\frac{\lambda_{1d}-\lambda_{0d}}{\lambda_{1d}\lambda_{0d}}\right). \quad (20)$$

The expected vector $\mathbf{M}_i$ and the covariance matrix $\mathbf{\Sigma}_i$ of $\boldsymbol{X}$, which is exponentially distributed according to (18), are given by $\mathbf{M}_i = (\lambda_{i1},\ldots,\lambda_{ij},\ldots,\lambda_{id})^T$ and $\mathbf{\Sigma}_i = \text{diag}(\lambda_{i1}^2,\ldots,\lambda_{ij}^2,\ldots,\lambda_{id}^2)$. Applying (12) to determine the projection direction under which the optimal value of any second-order criterion is reached, we obtain:

$$\boldsymbol{W}_\rho = (\frac{\lambda_{11}-\lambda_{01}}{\lambda_{11}^2-\rho(\lambda_{11}^2-\lambda_{01}^2)},\ldots,\frac{\lambda_{1j}-\lambda_{0j}}{\lambda_{1j}^2-\rho(\lambda_{1j}^2-\lambda_{0j}^2)}, \ldots,\frac{\lambda_{1d}-\lambda_{0d}}{\lambda_{1d}^2-\rho(\lambda_{1d}^2-\lambda_{0d}^2)}). \quad (21)$$

Comparing (20) and (21) shows that the collinearity of $\boldsymbol{W}^*$ and $\boldsymbol{W}_\rho$ depends on ρ. We shall now illustrate Scenarios 2 and 3.

Consider the case of two-dimensional observations $\boldsymbol{X}$ with $\lambda_{11}\neq\lambda_{01}$ and $\lambda_{12}\neq\lambda_{02}$. Vectors $\boldsymbol{W}^*$ and $\boldsymbol{W}_\rho$ are collinear if, and only if,

$$\rho = \frac{\lambda_{02}\lambda_{12}\lambda_{11}^2-\lambda_{01}\lambda_{11}\lambda_{11}^2}{\lambda_{02}\lambda_{12}(\lambda_{11}^2-\lambda_{01}^2)-\lambda_{01}\lambda_{11}(\lambda_{12}^2-\lambda_{02}^2)}. \quad (22)$$

This means that any second-order criterion Ψ guarantees the best solution in the Bayes sense if, and only if, it satisfies:

$$\frac{\frac{\partial\Psi}{\partial\sigma_0^2}}{\frac{\partial\Psi}{\partial\sigma_0^2}+\frac{\partial\Psi}{\partial\sigma_1^2}} = \frac{\lambda_{02}\lambda_{12}\lambda_{11}^2-\lambda_{01}\lambda_{11}\lambda_{11}^2}{\lambda_{02}\lambda_{12}(\lambda_{11}^2-\lambda_{01}^2)-\lambda_{01}\lambda_{11}(\lambda_{12}^2-\lambda_{02}^2)}, \quad (23)$$

an example of which is the generalized signal-to-noise ratio Ψ_α:

$$\Psi_\alpha(S) = \frac{(m_1-m_0)^2}{(1-\alpha)\sigma_1^2+\alpha\sigma_0^2}, \quad (24)$$

the parameter α must then be equal to the second term in (22). Here m_i and σ_i^2 denote the conditional expected values and variances of S. This illustrates Scenario 2.

Consider now that $\boldsymbol{X}$ is a d-dimensional observation, with ($d>2$). Except for very particular cases, we notice that there is no ρ parameter which ensures the collinearity of $\boldsymbol{W}^*$ and $\boldsymbol{W}_\rho$. This means that optimization of second-order criteria, relevant or not, does not necessarily lead to a Bayes-equivalent detector, even if there is one that is a member of $\mathcal{D}$. Figures 1, 2 and 3 illustrate this situation, called Scenario 3, for three-dimensional observations $\boldsymbol{X}$. The projection direction $\boldsymbol{W}_\rho$ is represented on the unit sphere as a function of $\rho\in]-\infty,+\infty[$. One can observe the collinearity of projection directions $\boldsymbol{W}_{-\infty}$ and $\boldsymbol{W}_{+\infty}$ (see Property 2). Finally, one can notice that there is no value ρ_0

such that $\boldsymbol{W}_{\rho_0}$ and $\boldsymbol{W}^*$ are collinear, which is one of the weaknesses of second-order criteria.

NONLINEAR SECOND-ORDER DISCRIMINANT

In recent years a great interest has been shown in kernel-based algorithms for developping a nonlinear generalization of linear receivers, see [11] and references therein. Kernel-based classification algorithms were primarily used in Support Vector Machines (SVMs) [3, 4]. By mapping the samples $(\boldsymbol{X}_i)_{i=1,\ldots,n}$ into a high dimensional feature space and reformulating the problem into dot product form in order to use Mercer kernels, an effective solution for nonlinear discriminant analysis has been obtained [17, chapter 5]. This exploits the notion that applying a nonlinear data transformation to some high-dimensional feature space increases the probability of having linearly separable classes in the transformed space. In [10], a nonlinear classification technique based on Fisher discriminants has been proposed. It also uses the Mercer kernel trick, and allows the efficient computation of linear Fisher discriminants in feature space. Very promising results had been reported using this approach, called the kernel Fisher discriminant method (KFD), when compared with other state of the art classification techniques. In this paper, we present an extension of the KFD method that also deals with nonlinear discriminant analysis using kernel functions and second-order measures of performance. This method is based on extending the expression (12) to the nonlinear case by mapping the data by a nonlinear transformation.

Formulation Of The Nonlinear Discriminant

Applying a nonlinear data transformation to some specific high dimensional feature spaces increases the probability of having linearly separable classes within the transformed space [17, 11]. Linear discriminants in the feature space are then equivalent to nonlinear discriminants in the original space. More formally, by mapping the samples $\{\boldsymbol{X}_i\}_{i=1,\ldots,n}$ using a nonlinear function

$$\begin{array}{rccc} \boldsymbol{\Phi}: & \mathbb{R}^d & \longrightarrow & \mathcal{F} \\ & \boldsymbol{X} & \longmapsto & \boldsymbol{\Phi}(\boldsymbol{X}), \end{array}$$

one can perform a linear discriminant analysis in $\mathcal{F}$ with the set $\{(\boldsymbol{\Phi}(\boldsymbol{X}_i), Y_i)\}_{i=1,\ldots,n}$ in order to obtain a nonlinear discriminants in the original space.

Clearly, if $\mathcal{F}$ is a very high, or even infinitely, dimensional space, deriving $S(\boldsymbol{X}) = (\boldsymbol{W})^T \boldsymbol{\Phi}(\boldsymbol{X}) - b$ may be a computationally intractable problem. However, by using the theory of reproducing kernels [3], such a problem can be solved without explicitly mapping the data to the feature space $\mathcal{F}$.

Nonlinear kernel second-order discriminant (KSOD) can be obtained by using (12) in the feature space $\mathcal{F}$. According to Proposition 1, the function $S(\boldsymbol{X})$ operating in $\mathcal{F}$ is optimum in the sense of any given second-order criterion Ψ if it satisfies

$$\left[\rho \boldsymbol{\Sigma}_0^{\boldsymbol{\Phi}} + (1-\rho)\boldsymbol{\Sigma}_1^{\boldsymbol{\Phi}}\right] \boldsymbol{W} = [\boldsymbol{M}_1^{\boldsymbol{\Phi}} - \boldsymbol{M}_0^{\boldsymbol{\Phi}}], \tag{25}$$

where $\boldsymbol{M}_i^{\boldsymbol{\Phi}}$ and $\boldsymbol{\Sigma}_i^{\boldsymbol{\Phi}}$ denote the conditional expected vectors and covariance matrices of $\boldsymbol{\Phi}(\boldsymbol{X})$, respectively. These moments can be estimated as follows:

$$\boldsymbol{M}_i^{\boldsymbol{\Phi}} = \frac{1}{n_i} \sum_{\boldsymbol{X} \in \mathcal{C}_i} \boldsymbol{\Phi}(\boldsymbol{X}) \tag{26}$$

$$\boldsymbol{\Sigma}_i^{\boldsymbol{\Phi}} = \frac{1}{n_i} \sum_{\boldsymbol{X} \in \mathcal{C}_i} \boldsymbol{\Phi}(\boldsymbol{X})\, \boldsymbol{\Phi}^T(\boldsymbol{X}) - (\boldsymbol{M}_i^{\boldsymbol{\Phi}})\, (\boldsymbol{M}_i^{\boldsymbol{\Phi}})^T, \tag{27}$$

where n_i is the number of samples from class $\mathcal{C}_i$ in the training set. When $\mathcal{F}$ is a very high-dimensional space, (25) may be difficult to solve except if the Mercer trick is used. From the theory of reproducing kernels [15], we know that any solution $\boldsymbol{W} \in \mathcal{F}$ must lie in the span of all training samples in $\mathcal{F}$. Therefore $\boldsymbol{W}$ can be written as follows:

$$\boldsymbol{W} = \sum_{i=1}^{n} \alpha(i)\, \boldsymbol{\Phi}(\boldsymbol{X}_i) = \mathbf{Q}\, \boldsymbol{\alpha} \tag{28}$$

where $\mathbf{Q}$ denotes the matrix $[\boldsymbol{\Phi}(\boldsymbol{X}_1) \cdots \boldsymbol{\Phi}(\boldsymbol{X}_n)]$, and the $\alpha(i)$'s are the dual parameters. Multiplying (25) by $\mathbf{Q}^T$ and using (28) yields

$$\left[\rho\, \mathbf{Q}^T \boldsymbol{\Sigma}_0^{\boldsymbol{\Phi}} \mathbf{Q} + (1-\rho)\, \mathbf{Q}^T \boldsymbol{\Sigma}_1^{\boldsymbol{\Phi}} \mathbf{Q}\right] \boldsymbol{\alpha} = \mathbf{Q}^T [\boldsymbol{M}_1^{\boldsymbol{\Phi}} - \boldsymbol{M}_0^{\boldsymbol{\Phi}}]. \tag{29}$$

Let k be any kernel that satisfies the Mercer condition [17, 11]. Then we have

$$k(\boldsymbol{X}_i, \boldsymbol{X}_j) = \boldsymbol{\Phi}(\boldsymbol{X}_i)^T \boldsymbol{\Phi}(\boldsymbol{X}_j), \tag{30}$$

which means that $k(\boldsymbol{X}_i, \boldsymbol{X}_j)$ corresponds to the inner product of $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ in $\mathcal{F}$.

By using a kernel which verifies (30), the expression (29) can be reformulated as

$$\mathbf{P}_\rho\, \boldsymbol{\alpha} = \boldsymbol{M}, \tag{31}$$

where $\boldsymbol{\alpha}$ has to be determined and $\mathbf{P}_\rho$ is a n by n matrix which is given by

$$\mathbf{P}_\rho = \left[\frac{\rho}{n_0}\, \mathbf{K}_0 (\mathbf{I} - \mathbf{1}_{n_0}) \mathbf{K}_0^T + \frac{1-\rho}{n_1}\, \mathbf{K}_1 (\mathbf{I} - \mathbf{1}_{n_1}) \mathbf{K}_1^T\right]. \tag{32}$$

In the above expression, $\mathbf{K}_i$ is a n by n_i matrix with elements

$$\mathbf{K}_i(p, q) = k(\boldsymbol{X}_p, \boldsymbol{X}_q), \tag{33}$$

for all $\boldsymbol{X}_p \in (\mathcal{C}_0 \cup \mathcal{C}_1)$ and $\boldsymbol{X}_q \in \mathcal{C}_i$. $\mathbf{I}$ is the identity matrix and $\mathbf{1}_{n_i}$ is the matrix with all elements set to $\frac{1}{n_i}$. The components of $\boldsymbol{M}$ in (31) are defined as

$$M(j) = \frac{1}{n_1} \sum_{\boldsymbol{X} \in \mathcal{C}_1} k(\boldsymbol{X}, \boldsymbol{X}_j) - \frac{1}{n_0} \sum_{\boldsymbol{X} \in \mathcal{C}_0} k(\boldsymbol{X}, \boldsymbol{X}_j). \quad (34)$$

To determine the projection of any new sample $\boldsymbol{X}$ onto $\boldsymbol{W}$, we have to calculate the n-dimensional vector $\boldsymbol{\alpha}$ from (31). Eq. (28) yields:

$$\boldsymbol{\Phi}(\boldsymbol{X})^T \boldsymbol{W} = \sum_{i=1}^{n} \alpha(i)\, k(\boldsymbol{X}_i, \boldsymbol{X}). \quad (35)$$

Finally one can use different strategies to determine the bias b, as shown in [10]. We can summarize our work in the algorithm described below to determine a minimum error rate detector from a class of kernel-based decision structures that are optimum in the sense of second-order criteria.

1. Given any Mercer kernel k, compute the kernel matrices $\mathbf{K}_0$ and $\mathbf{K}_1$ from (33).
2. Compute $\boldsymbol{M}$ from (34).
3. Set ρ to zero.
4. While ($\rho \leq 1$) repeat
   - Compute $\mathbf{P}_\rho$ from (32).
   - Solve (31) to get the vector $\boldsymbol{\alpha}$, and retain the result as $\boldsymbol{\alpha}_\rho$.
   - Find the threshold b_ρ which minimizes, e.g., an estimate of the generalization error.[2]
   - Update ρ: $\rho \leftarrow \rho + \Delta\rho$, where $\Delta\rho$ is a selected step.
5. Select the best detector characterized by $(\boldsymbol{\alpha}_\rho, b_\rho)$.

From equation (13), one can see that the parameter ρ varies from 0 to 1 if the derivatives of the corresponding second-order criteria Ψ with respect to σ_0^2 and σ_1^2 have the same sign. This desirable but non-mandatory requirement for design criteria makes the process much simpler than adjusting ρ in $\mathbb{R}$.

The method presented above is called *kernel second-order discriminant* since it leads to the optimum nonlinear receiver in the sense of the best second-order criterion without setting it up. Obviously, classifiers obtained with KSOD perform better than or equal to those resulting from the KFD method developed by Mika *et al.* in [10].

[2] See, e.g., [10] for other criteria and strategies.

Experimentations

To compare our method to KFD and SVM, 10 experiments were conducted on artificial and real world data downloaded from `http://www.first.gmd.de/~raetsch`. For each of the 10 problems, Table 1 shows the average test error over 40 runs on 400 training samples and 8000 test samples chosen arbitrarily from a mixture of the available data sets. The kernel function was selected as the RBF having a width equal to 1 [17]. Note that the numerical problems created by inverting the ill-conditioned matrix $\mathbf{P}$ in (31) can be avoided by adding a regularization term. It can be chosen as a multiple of the identity matrix [10, 11, 17], i.e., replace $\mathbf{P}_\rho$ by the matrix $\mathbf{P}_\rho + \eta\mathbf{I}$ with $\eta > 0$. The results presented in Table 1 clearly show that the KSOD method can often perform favourably compared with the other state of the art detection techniques. However, note that for the KSOD method, all the training data should be used to test a new sample. This is not the case for the SVM method which uses only the training samples called support vectors [17]. Hence, the testing time for the KSOD method is in general higher than that of the SVMs. Table 1 gives the number of support vectors used by the SVM method.

CONCLUSION

The theoretical results reported in this paper are concerned with the virtues and vices of second-order criteria used for detector design. Firstly, we have given a necessary and sufficient condition for these measures of performance to guarantee the best solution in the Bayes sense when deriving unconstrained detectors. Secondly we have considered the case where constraints are imposed on the structure of detectors, leading us to restrict our attention to a class $\mathcal{D}$ of admissible detectors. We have shown that any second-order criterion, relevant or not, can sometimes lead to a Bayes-optimal receiver. However, such a success is far from assured in the majority of cases, even if a Bayes-equivalent detector in $\mathcal{D}$ exists.

Finally, we have proposed a nonlinear classification technique based on second-order criteria. It uses the Mercer kernel trick, and allows for the efficient computation of linear second-order discriminants in feature space. The results obtained suggest that the method presented in this paper often performs better than SVM and KFD. However, one should note that the testing time for this method is in general higher than that of the SVMs since it uses all the training samples in the test phase. Future works will be dedicated to control the complexity of the resulting discriminant

in order to increment the generalization performances and to reduce the testing time [11, 16, 9].

REFERENCES

[1] F. Abdallah, C. Richard and R. Lengell. On equivalence between detectors obtained from second-order measures of performance. *XI European Signal Processing Conference*, September 3-6, 2002, Toulouse, France.

[2] S. A. Billings and K. L. Lee. "Nonlinear Fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm," *Neural Networks*, vol. 15, pp. 263-270, 2002.

[3] B. Boser, I. Guyon and V. Vapnik. "A training algorithm for optimal margin classifiers," *Proc. Fifth Annual Workshop on Computational Learning Theory*, Pittsburgh, pp. 144-152, 1992.

[4] C. Cortes and V. Vapnik. "Support vector networks," *Machine learning* , vol. 20, pp. 273-297, 1995.

[5] L. Devroye, L. Gyorfi and G. Lugosi. *A probabilistic theory of pattern recognition*, Springer-Verlag, 1996.

[6] R. O. Duda, P. E. Hart and D. C. Stork. *Pattern classification*. New York: John Wiley and Sons, 2001.

[7] K. Fukunaga. *Statistical Pattern Recognition*. Academic Press, 1990.

[8] W. A. Gardner. A unifying view of second-order measures of quality for signal classification. *IEEE Transactions on Communications*, vol. 28, no. 6, pp. 807–816, 1980.

[9] Y. LeCun, J. S. Denker and S. A. Solla. "Optimal brain damage," in *Proc. Advances in Neural Information Processing Systems*, vol. 2, pp. 598–605, 1990.

[10] S. Mika, G. Rätsch, J. Weston, B. Schölkopf and K. R. Müller. Fisher discriminant analysis with kernels, in *Advances in Neural Networks for Signal Processing* , Y. H. Hu, J. Larsen, E. Wilson, S. Douglas, editors, pp. 41-48, 1999.

[11] K. R. Müller, S. Mika, G. Rätsch, K. Tsuda and B. Schölkopf. "An introduction to kernel-based learning algorithms," *IEEE Neural Networks*, vol. 12, no. 2, pp. 181-201, May 2001.

[12] D. W. Patterson and R. L. Mattson. A method of finding linear discriminant functions for a class of performance criteria. *IEEE Transactions on Information Theory*, vol. 12, no. 3, pp. 380-387, 1966.

[13] H. V. Poor. *An introduction to signal detection and estimation*. Springer-Verlag, 1994.

[14] C. Richard, R. Lengell and F. Abdallah. Bayes-optimal detectors design using relevant second-order criteria. *IEEE Signal Processing Letters*, vol. 9, no. 1, 2002.

[15] S. Saitoh. *Theory of Reproducing Kernels and its Applications*. Longman Scientific & Technical, 1988.

[16] J. Rissanen. "Stochastic complexity in statistical inquiry," *Annals of Statistics*, vol. 14, no. 3, pp. 1080–1100, 1986.

[17] V. Vapnik.*The nature of statistical learning theory*. New York: Springer Verlag, 1995.

NOTATIONS

All vectors and matrices are in bold	
$\mathcal{D}$	class of functions or detectors
$\mathcal{C}_i$	class i
$\boldsymbol{X}$	input sample
$X(i)$	the component number i of $\boldsymbol{X}$
Y_i	target value for class i
d	dimensionality
n	number of training samples
$\mathcal{F}$	high dimensional feature space
$p_i(\boldsymbol{X})$	conditional probability density of class i
$S(.)$	mesurable function
$g(.)$	decision function
$L(.)$	likelihood ratio
Ψ	second order criterion
m_i	expected value with respect to class i
σ_i	variance with respect to class i
$\boldsymbol{W}$	d-dimensional vector
$\boldsymbol{\alpha}$	n dimensional dual vector
$\alpha(i)$	the component number i of $\boldsymbol{\alpha}$
$k(.,.)$	Mercer kernel
$\boldsymbol{\Phi}$	nonlinear function
$\boldsymbol{M}_i$	conditional expected vector of class i
$\boldsymbol{M}_i^{\boldsymbol{\Phi}}$	conditional expected vector of class i in feature space
$\boldsymbol{\Sigma}_i$	conditional covariance matrix of class i
$\boldsymbol{\Sigma}_i^{\boldsymbol{\Phi}}$	conditional covariance matrix of class i in feature space
ROC	Receiver Operating Characteristic
KFD	Kernel Fisher Detectors
SVM	Support Vector Machine
KSOD	Kernel Second Order Detectors