On virtues and vices of second-order measures of quality for binary classification

Fahed Abdallah, Cédric Richard, Régis Lengellé

Laboratoire de Modélisation et Sûreté des Systèmes (LM2S),

Université de Technologie de Troyes (UTT),

12 rue Marie Curie, B.P. 2060,

10010 Troyes cedex, France.

tel: +33.3.25.71.56.92 fax: +33.3.25.71.56.99

E-mail: fahed.abdallah@utt.fr

Key words liste I: Pattern recognition

Key words liste II: Maximum likelihood classifier, Bayes error, Signal-to-noise ratio, Classifiers, Classification, Detection, Distance measures, Decision making, Discriminate analysis, Feature detection, Statistics, Signal processing, Performance analysis, Probability density function, Gaussian

Abstract

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, and thus 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 $\mathcal{C}$ of admissible detectors. In these conditions we prove that even if it exists a monotonic function of the likelihood ratio in $\mathcal{C}$, obtaining this statistic via the optimization of a second-order criterion, relevant or not, is not guaranteed. Finally, results are illustrated with simulation examples.

1 Introduction

Let (X, Y) be a pair of random variables taking their respective values from $\mathbb{R}^n$ and $\{0, 1\}$, where X is the observation and Y is the class. The purpose of detection is to determine to which of two classes $Y = 0$ or $Y = 1$ a given observation X belongs. According to classical statistical detection theories, comparing any strictly monotonic function of the likelihood ratio $L(X)$ with a threshold value is the optimum test [Poor, 1994]. In practical applications, implementing such a test may be impossible because of incomplete specification of the conditional probability densities $p(X|Y = 0)$ and $p(X|Y = 1)$, denoted by the standard notations $p_0(X)$ and $p_1(X)$, 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 [Duda and Hart, 1973, Fukunaga, 1990]. A wide variety of second-order measures of performance have been proposed and several contributions have been presented to prove their efficiency, e.g., [Gardner, 1980] 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 crterion, mean-square error and signal-to-noise ratio.

The aim of this paper is to provide an overview of virtues and vices 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, and thus 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 $\mathcal{C}$ of admissible detectors. In these conditions, we prove that even if it exists a monotonic function of the likelihood ratio in $\mathcal{C}$, obtaining this statistic via the optimization of a relevant criterion is not guaranteed. Finally, results are illustrated with simulations.

2 Characterization of relevant second-order criteria

2.1 Background and notations

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

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

which errs on X if $g(X) \neq Y$. Bayes decision theory leads to the result of major importance that the optimum detector g^* in the sense that it minimizes the risk consists in comparing the likelihood ratio $L(X) \triangleq p_1(X)/p_0(X)$ with a given threshold ν in order to make a decision [Poor, 1994]. This decision rule can be expressed as

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

Note that g is equivalent[1] to the Bayes-optimal detector g^* if $S(X) = \Phi\{L(X)\}$, 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 $S(X)$, 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 explore, individually, the relevancy of well-known second-order criteria such as Fisher, mean-square error and signal-to-noise ratio (see, e.g., [Gardner, 1980] and references therein). In [Fukunaga, 1990, 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 provide Bayes-optimal detectors, and thus to be considered as a *relevant second-order criterion* for detector design.

2.2 Relevant second-order criteria

Let Ψ be any function of $m_i \triangleq \int S(X)\, p_i(X)\, dX$ and $\sigma_i^2 \triangleq \int (S(X) - m_i)^2\, p_i(X)\, dX$, where $S(X)$ denotes any decision statistic. We first have to characterize statistics $S(X)$ which optimize Ψ. Operating on Ψ

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

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

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

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

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

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

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

We thus note that $S(X)$ defined by (6) is a strictly monotonic function of $L(X)$ if, and only if, the numerator of (7) is not equal to 0. This result leads directly to the following proposition [Richard et al, 2002].

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 [Abdallah et al, 2002]. This result is one of the most interesting virtues of second-order criteria.

3 Constrained detector design using second-order criteria

As shown in Section 2, there exists a broad class of second-order criteria that lead to a detection statistic $S(X)$ equivalent to the likelihood ratio $L(X)$. However, implementing $S(X)$ remains an unsolved problem since it depends on the probability densities $p_0(X)$ and $p_1(X)$ via $L(X)$, which are unknown. Therefore, we are often led to consider the following strategy for deriving receivers [Devroye et al, 1996]:

1. selecting a class $\mathcal{C}$ of detection statistics;
2. picking the statistic of $\mathcal{C}$ 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 $\mathcal{C}$. We shall now discuss this drawback in the case where $\mathcal{C}$ denotes the class of linear statistics.

3.1 Linear detectors design

Linear classifiers are the simplest one as far as implementation is concerned, and are directly related to many known techniques such as correlations and Euclidean distances [Fukunaga, 1990]. We shall now show how second-order criteria can be used for designing linear detectors, which are defined as follows:

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

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

$$m_i = \mathrm{E}\{S \,|\, Y = i\} = K^T M_i - \nu \tag{10}$$

$$\sigma_i^2 = \mathrm{Var}\{S \,|\, Y = i\} = K^T \Sigma_i K, \tag{11}$$

where M_i and Σ_i are the conditional expected vectors and covariance matrices of X. Let Ψ be any second-order criterion. The optimal statistic $S(X)$ is given by equating to zero the partial derivatives of Ψ with respect to K and ν. As shown in [Fukunaga, 1990, pp. 133-4] and [Patterson and Mattson, 1966], solving this linear system leads directly to the following proposition.

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

$$K_\rho = [\rho \Sigma_0 + (1-\rho)\Sigma_1]^{-1} [M_1 - M_0], \tag{12}$$

where M_i and Σ_i are the conditional expected vectors and covariance matrices of 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 K_ρ depends on Ψ through a single parameter $\rho \in]-\infty, +\infty[$. Then the latter can chosen so as 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 $K_{-\infty}$ and $K_{+\infty}$. They are both proportional to $[\Sigma_0 - \Sigma_1]^{-1}[M_1 - M_0]$ since we have

$$K_{\pm\infty} \propto \lim_{\rho \to \pm\infty} \frac{1}{\rho}[\Sigma_0 - \Sigma_1]^{-1}[M_1 - M_0]. \tag{14}$$

up to a scaling factor such that $\|K_{\pm\infty}\| = 1$. The projection directions $K_{-\infty}$ and $K_{+\infty}$ then lead to equivalent detection structures (**Property 2**).

In the following, Proposition 2 and the above properties are illustrated through some classic detection problems. As mentioned at the very beginning of Section 3, we also show that (relevant) second-order criteria does not guaranty an optimal detector in the Bayes sense if we restrict the solution space to a specific class $\mathcal{C}$ of detectors. Here this drawback is illustrated through the following situations:

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

Scenario 2: there exist second-order criteria that provide Bayes-optimal receivers in $\mathcal{C}$,

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

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

3.2 Scenario 1: case of normal distributions with equal covariances

When $p_0(X)$ and $p_1(X)$ are normal with expected vectors M_0 and M_1 and covariance matrices Σ_0 and Σ_1, it is well-known that the Bayes-optimal statistic is given by:

$$S(X) = \frac{1}{2}(X - M_0)^T \Sigma_0^{-1}(X - M_0) - \frac{1}{2}(X - M_1)^T \Sigma_1^{-1}(X - M_1). \tag{15}$$

This equation shows that the decision boundary is a quadratic form in X. When $\Sigma_0 = \Sigma_1 = \Sigma$, the boundary becomes a linear function of X as

$$S(X) = (M_1 - M_0)^T \Sigma^{-1} X. \tag{16}$$

Eq. (16) indicates that the direction onto which any X is projected is given by $K^* = \Sigma^{-1}(M_1 - M_0)$. Let us now determine the projection direction under which the optimal value of any second-order criterion is reached. Eq. (12) gives:

$$K_\rho = \Sigma^{-1}(M_1 - M_0). \tag{17}$$

Comparing the projection directions K^* and K_ρ, 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 $\mathcal{C}$.

3.3 Scenarios 2 and 3: case of exponential distributions

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

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

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:

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

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

$$K^* = \left(\frac{\lambda_{11} - \lambda_{01}}{\lambda_{11}\lambda_{01}}, \ldots, \frac{\lambda_{1j} - \lambda_{0j}}{\lambda_{1j}\lambda_{0j}}, \ldots, \frac{\lambda_{1n} - \lambda_{0n}}{\lambda_{1n}\lambda_{0n}}\right). \tag{20}$$

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

$$K_\rho = \left(\frac{\lambda_{11} - \lambda_{01}}{\lambda_{11}^2 - \rho(\lambda_{11}^2 - \lambda_{01}^2)}, \ldots, \frac{\lambda_{1j} - \lambda_{0j}}{\lambda_{1j}^2 - \rho(\lambda_{1j}^2 - \lambda_{0j}^2)}, \ldots, \frac{\lambda_{1n} - \lambda_{0n}}{\lambda_{1n}^2 - \rho(\lambda_{1n}^2 - \lambda_{0n}^2)}\right). \tag{21}$$

Comparing (20) and (21) shows that the collinearity of K^* and K_ρ depends on ρ. We shall now illustrate Scenarios 2 and 3 described in Section 3.1.

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

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

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

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

an example of which is the generalized signal-to-noise ratio Ψ_α with α given by (22):

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

Here m_i and σ_i^2 denote the conditional expected values and variances of S. This illustrates Scenario 2.

Consider now that X is a n-dimensional observation, with $(n > 2)$. Except for very particular cases, we can notice that it does not exist any ρ ensuring collinearity of K^* and K_ρ. 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 $\mathcal{C}$. Figures 1, 2 and 3 illustrate this situation, called Scenario 3 in Section 3.1, for three-dimensional observations X. The projection direction K_ρ is represented on the unit sphere as a function of $\rho \in]-\infty, +\infty[$. One can observe the collinearity of projection directions $K_{-\infty}$ and $K_{+\infty}$ (see Property 2, Section 3.1). Finally, one can notice that there is no value ρ_0 such that K_{ρ_0} and K^* are collinear, which is one of the weaknesses of second-order criteria.

4 Conclusion

The theoretical results reported in this paper are concerned with the virtues and vices of second-order criteria used for detector design. First, we have given a necessary and sufficient condition for these measures of performance to guaranty the best solution in the Bayes sense when deriving unconstrained detectors. Next we have considered the case where constrained are imposed on the structure of detectors, leading us to restrict our attention to a class $\mathcal{C}$ 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 general case, even if exists a Bayes-equivalent detector in $\mathcal{C}$.

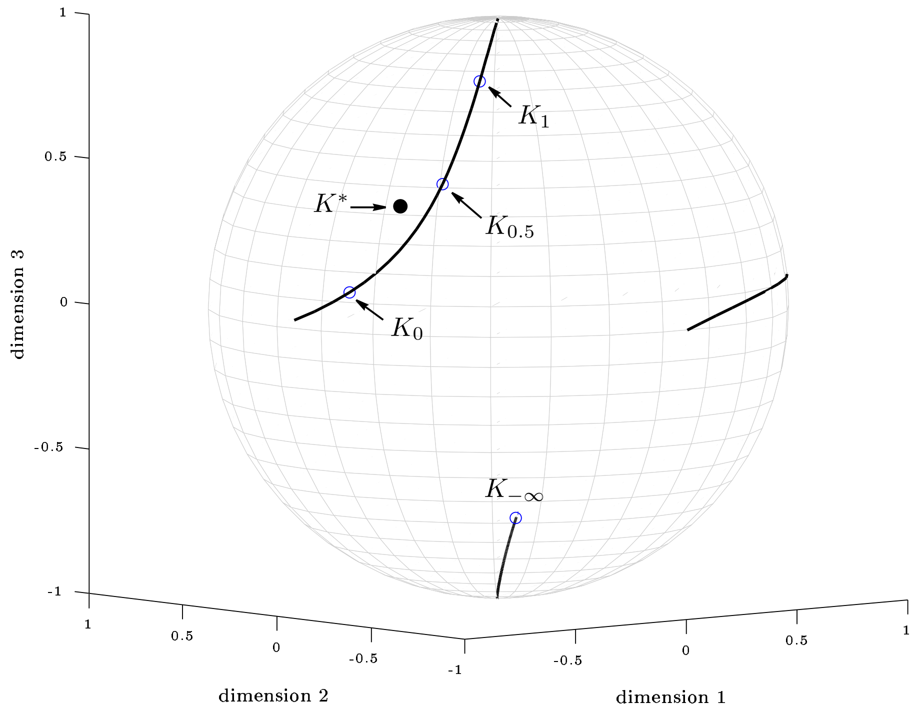

Figure 1: Evolution of the projection direction K_ρ on the unit sphere as a function of $\rho \in]-\infty, +\infty[$, in the case of three-dimensional exponential distributions ($\lambda_{01} = 5$, $\lambda_{11} = 2$, $\lambda_{02} = 3$, $\lambda_{12} = 2$, $\lambda_{03} = 2$, $\lambda_{13} = 3$). The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

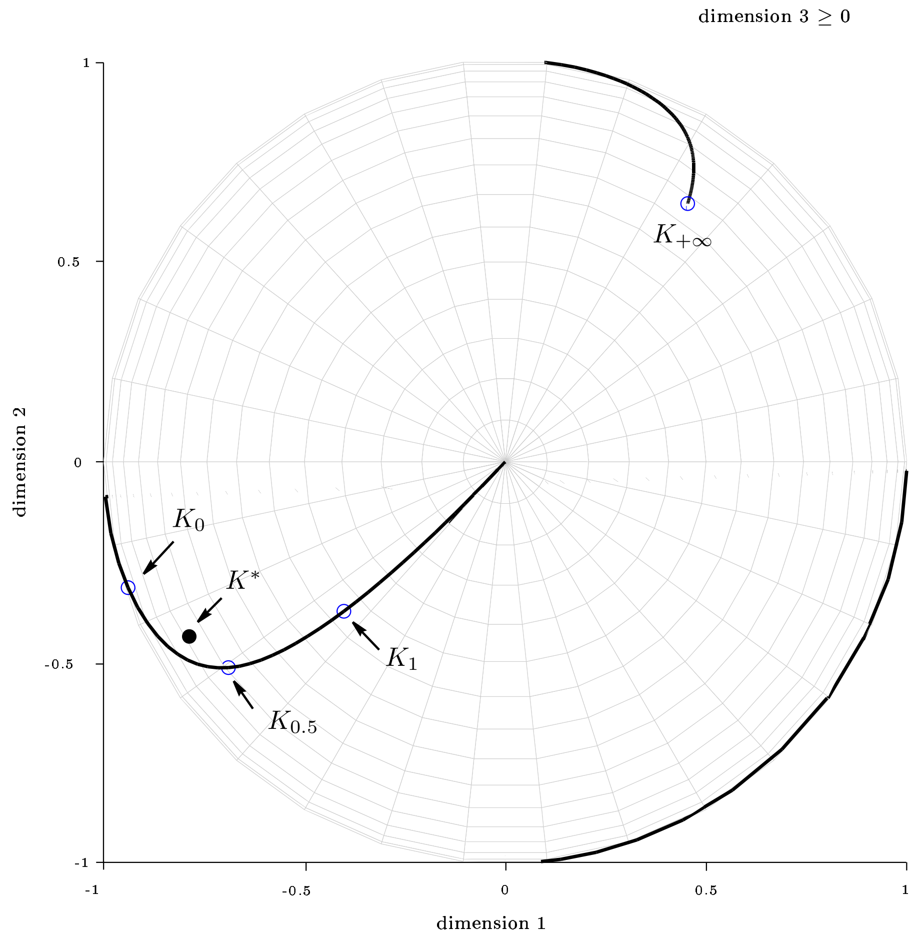

Figure 2: Same as figure 1. Overhead view of the unit sphere, i.e., dimension $3 \geq 0$.

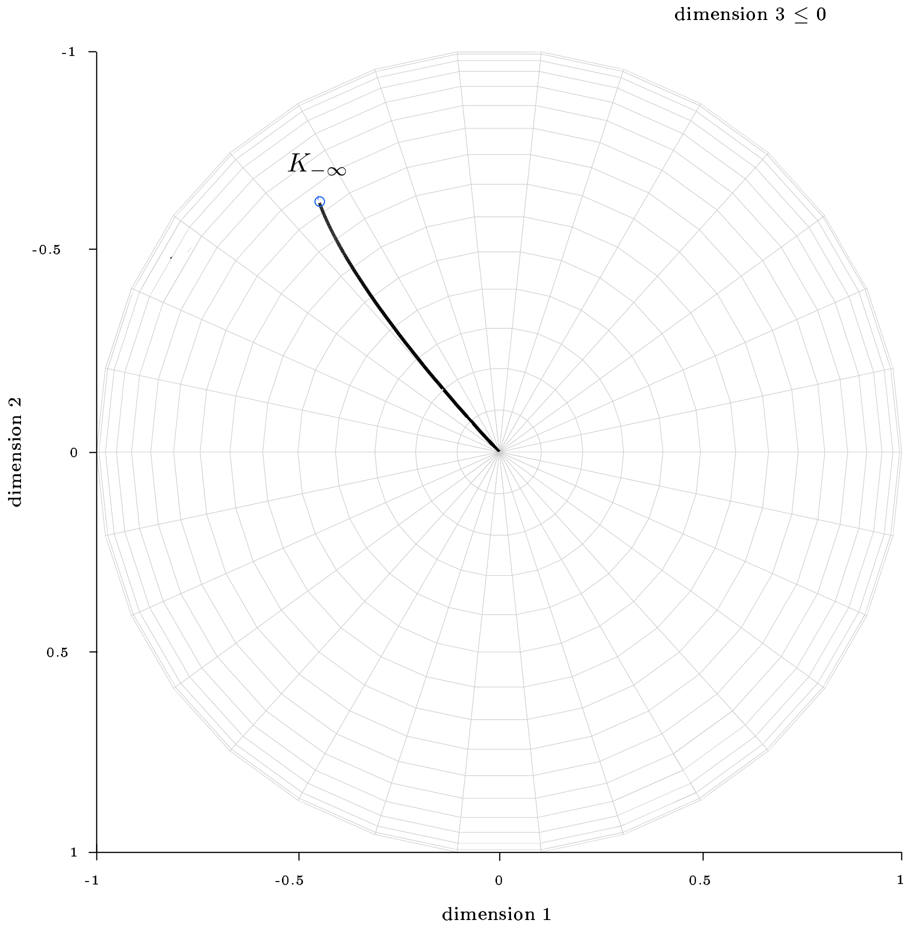

Figure 3: Same as figure 1. View from below of the unit sphere, i.e., dimension 3 ≤ 0.

References

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

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

[Duda and Hart, 1973] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*, Wiley and Sons, 1973.

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

[Gardner, 1980] 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.

[Patterson and Mattson, 1966] 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.

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

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