Annals of Operations Research
https://doi.org/10.1007/s10479-019-03344-6

S.I.: STOCHASTIC MODELING AND OPTIMIZATION, IN MEMORY OF ANDRÁS PRÉKOPA

Risk-averse classification

Constantine Alexander Vitt[1] · **Darinka Dentcheva**[2] · **Hui Xiong**[1]

Abstract
We develop a new approach to solving classification problems, which is based on the theory of coherent measures of risk and risk sharing ideas. We introduce the notion of a risk-averse classifier and a family of risk-averse classification problems. We show that risk-averse classifiers are associated with minimal points of the possible classification errors, where the minimality is understood with respect to a suitable stochastic order. The new approach allows for measuring risk by distinct risk functional for each class. We analyze the structure of the new classification problem and establish its theoretical relation to known risk-neutral design problems. In particular, we show that the risk-sharing classification problem is equivalent to an implicitly defined optimization problem with unequal weights for each data point. Additionally, we derive a confidence interval for the total risk of a risk-averse classifier. We implement our methodology in a binary classification scenario on several different data sets. We formulate specific risk-averse support vector machines in order to demonstrate the proposed approach and carry out numerical comparison with classifiers which are obtained using the Huber loss function and other loss functions known in the literature.

Keywords Machine learning · Support vector machines · Soft-margin classifier · Coherent measures of risk · Risk sharing · Normalized classifiers · Risk-aware classification

1 Introduction

Classification is one of the fundamental tasks of the data mining and machine learning community. The need for accurate and effective solution of classification problems prolif-

This paper is dedicated to András Prékopa in recognition of his fundamental contributions to probability and to optimization under uncertainty.

✉ Darinka Dentcheva
darinka.dentcheva@stevens.edu

Constantine Alexander Vitt
constantine.vitt@rutgers.edu

Hui Xiong
hxiong@rutgers.edu

[1] Rutgers University, Newark, New Brunswick, NJ, USA

[2] Stevens Institute of Technology, Hoboken, NJ, USA

Published online: 09 August 2019

erates throughout the business world, engineering, and sciences. In this paper, we propose a new approach to classification problems with the aim to develop a methodology for reliable and robust risk-averse classifier design which has allows the users to choose tailored risk measurement for misclassification in various classes. Classification problems are based on observed data; they use approximations for the true distribution of the populations to be separated. Creating a good approximation or taking into account the uncertainty of the approximated distribution is important for drawing proper conclusion. The uncertainty is exacerbated when the data is high dimensional but some or all populations are represented by small (relative to the dimensionality) samples. Furthermore, we stipulate that misclassification in different classes is associated with different risk. Naturally, when the sample sizes of the populations are imbalanced, the the statistical estimates associated with the small size samples carry more risk than those based on large samples. Our approach contributes to the methods for classification of imbalanced classes. Outside of that scenario, misclassification for different classes may be associated with dramatically different cost which should be taken into account when designing a classifier. We comment further on that issue in due course.

The proposed approach has its foundation in the theory of coherent measures of risk and risk sharing. Although, this theory is well advanced in the field of mathematical finance and actuarial analysis, the classification problem does not fit the problem setting analyzed in those fields and the theoretical results on risk sharing are inapplicable here. The classification problem raises new issues, poses new challenges, and requires a dedicated analysis. We employ non-linear in probability risk functionals specific to each class. We analyze the structure of the new classifier design problem and establish its theoretical relation to the risk-neutral design problem. In particular, we show that the risk-sharing classification problem is equivalent to an implicitly defined optimization problem with unequal, implicitly defined but unknown weights for each data point. We implement our methodology in a binary classification scenario on several different data sets and carry out numerical comparison with classifiers which are obtained using the Huber loss function and other popular loss functions. In these applications, we use linear support vector machines in order to demonstrate the proposed approach.

Our paper is organized as follows. In Sect. 2, we introduce the loss function and provide a formal definition to the problem illustrating it by examples. In Sect. 3, we introduce the necessary notions; we recall the notion of law-invariant coherent measures of risk and their dual representation, which illustrates how those risk measures provide robustness to the solution of a stochastic optimization problem. The main results of our paper are contained in Sect. 4. In Sect. 5, we derive confidence intervals for the misclassification risk as measured by the coherent measures of risk. In Sect. 6, we address specifically risk-averse binary classification by proposing risk-averse problem formulation, which can be solved in an efficient way numerically. In Sect. 7, we refer to some related work in the area of robust statistics, robust optimization, and measures of risk. Additionally, we discuss the problem of risk sharing and risk allocation in financial institutions. Finally, Sect. 8 reports on our numerical experiments.

2 Problem setting

We consider labeled data consisting of k subsets $S_1, \ldots, S_k$ of n-dimensional vectors. The cardinality of S_i is $|S_i| = m_i, i = 1, \ldots, k$. Analytically, the classification problem consists

of identifying a mapping ϕ, whose image is partitioned into k subsets corresponding to each class of data, so that $\phi(\cdot)$ can be used as an indicator function of each class. We adopt the following definition.

Definition 1 A classifier is a vector function $\varphi : \mathbb{R}^n \to \mathbb{R}^d$ such that $\varphi(x) \in K_i$ for all $x \in S_i$, $i = 1, \ldots, k$, where $K_i \subset \mathbb{R}^d$ and $K_i \cap K_j = \emptyset$ for all $i, j = 1, \ldots, k$ and $i \neq j$.

In our discussion, we assume that the classifier belongs to a certain functional family depending on a finite number of parameters, which we denote by $\pi \in \mathbb{R}^s$. The task is to choose a suitable value for the parameter π.

Some examples of this point of view are the following. When support vector machine is formulated, we seek to distinguish two classes, i.e., $k = 2$. The classifier is a linear function $\varphi(x; \pi) : \mathbb{R}^n \to \mathbb{R}$, defined by setting

$$\varphi(x; \pi) = v^\top x - \gamma \quad \text{for any } \ x \in \mathbb{R}^n.$$

The classifier is determined by the parameters $\pi = (v, \gamma) \in \mathbb{R}^{n+1}$. The regions the classifier maps to are $K_1 = [0, +\infty)$, $K_2 = (-\infty, 0)$.

Another example is given, when we wish to separate many classes ($k \geq 3$) by a linear classifier, which is created on the principle "one vs. all". Then effectively, our goal is to determine functions $\varphi_j(x; a^i, b_i) := \langle a^i, x\rangle - b_i$, where x is a data point, $a^i \in \mathbb{R}^n$, $i = 1, \ldots k-1$, are the normals of the separating planes and b_i determine the location of the i-th plane. Plane i is meant to separate the data points from class j from the rest of the data points. This means that

$$\varphi_j(x; a^i, b_i) = \begin{cases} \geq 0 & \text{for } x \in S_i \\ < 0 & \text{for } x \notin S_i. \end{cases} \tag{1}$$

We define a $k-1 \times n$ matrix A whose rows are the vectors a^i, and a vector $b \in \mathbb{R}^{k-1}$ whose components are b_i. The classifier for this problem can be viewed as a vector function $\varphi(\cdot; A, b) : \mathbb{R}^n \to \mathbb{R}^{k-1}$ by setting $\varphi(x; A, b) = Ax - b$. The parameter space is of form $\pi = (A, b) \in \mathbb{R}^{(k-1)(n+1)}$. Requirement (1) means that the regions K_j are the orthants

$$K_i = \{z \in \mathbb{R}^{k-1} : z_i \geq 0,\ z_j < 0, \quad j \neq i,\ j = 1, \ldots, k-1\},\ i = 1, \ldots k-1;$$
$$K_k = \{z \in \mathbb{R}^{k-1} : z_i < 0,\ i = 1, \ldots, k-1\}$$

This setting may be used for classification in some anomaly detection scenarios. Two approaches are known. One setting may require to distinguish between several distinct normal regimes or features of normal operational status. In that case, the class k may contain the anomalous instances, while classes $i = 1, \ldots k-1$ represent the normal operation. Another problem deals with several rare undesirable phenomena with distinct features. In such a scenario, we may associate classes $i = 1, \ldots k-1$ with those anomalous events and class k with a normal operation. When kernels are used, then the mapping $\varphi(x; \pi)$ becomes a composition of the embedding mapping from $\mathbb{R}^n$ to the new feature space and another classifier mapping in the feature space.

For a random observation $z \in \mathbb{R}^n$, we calculate $\varphi(z; \pi)$ and note that misclassification occurs when $\varphi(z; \pi) \notin K_i$, while $z \in S_i$ for any $i = 1, \ldots, k$. The classification error can be defined as the distance of a particular record to the classification set, to which it should belong. This definition is in harmony with the notion of an error in statistics, when a model is fit to data, in which case the error is defined as the distance of the model prediction to the

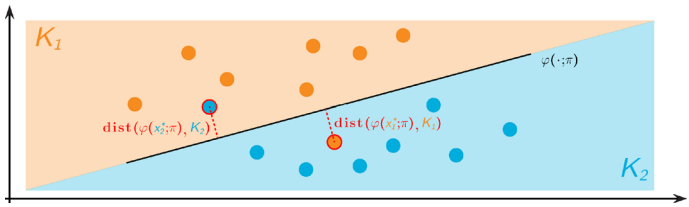

Fig. 1 Classification error calculation

realizations of the predicted random variable. Here the distance from a point r to a set K is defined by using a suitable norm in $\mathbb{R}^n$:

$$\mathrm{dist}(r, K) = \min\{\|r - a\| : a \in K\}.$$

The distance is well-defined when the set K is convex and closed.

As the records in every data class S_i, $i = 1, \ldots, k$ constitute a sample of an unknown distribution of a random vector X^i defined on a probability space $(\Omega, \mathcal{F}, P)$, the following random variables:

$$Z^i(\pi) = \mathrm{dist}(\varphi(X^i; \pi), K_i), \quad i = 1, \ldots k, \tag{2}$$

represent the (random) misclassification of data points in class i when parameter π is used. These univariate random variables are defined on a common probability space and are represented by the sampled observations

$$z_j^i(\pi) = \mathrm{dist}(\varphi(x_j^i; \pi), K_i) \quad \text{with} \quad x_j^i \in S_i \quad j = 1, \ldots, m_i.$$

The distance of $\varphi(x_j^i; \pi)$ to K_i is the smallest translation needed to eliminate misclassification of the point x_j^i. Figure 1 illustrates how the classification error for a certain binary classifier is measured. In the support vector machine, the classification error is computed by

$$\mathrm{dist}\big(\varphi(x; v, \gamma), K_i\big) = \begin{cases} \max(0, \langle v, x\rangle - \gamma) & \text{for } x \in S_1, \\ \max(0, \gamma - \langle v, x\rangle) & \text{for } x \in S_2. \end{cases}$$

We classify every new observation x in S_i, if $\mathrm{dist}\big(\varphi(x; v, \gamma), K_i\big) = 0, i, = 1, 2$. In the case of SVM, the regions cover the entire image space of the classifier $R = K_1 \cup K_2$. Therefore, the condition $\mathrm{dist}\big(\varphi(x; v, \gamma), K_i\big) = 0, i, = 1, 2$, always holds for exactly one class.

Observe that in the multi-class example, the regions $K_i, i = 1, \ldots k$ do not cover the entire image space of the classifier. Therefore, it is possible to observe a future instance x such that $\mathrm{dist}\big(\varphi(x; A, b), K_i\big) > 0$ for all $i = 1, \ldots k$. In that case, we could classify according to the smallest distance

$$x \in S_j \quad \text{iff} \quad \mathrm{dist}\big(\varphi(x; A, b), K_j\big) = \min_{1 \le i \le k} \mathrm{dist}\big(\varphi(x; A, b), K_i\big), \quad j \in \{1, \ldots, k\}.$$

Another problem arises, if the the minimum distance is achieved for several classes. The ambiguity could be resolved in several ways as a sequential classification procedure but this question is beyond the scope of our study.

If the distribution of the vectors $X^i, i = 1, \ldots, k$, are known, then the optimal risk-neutral classifier would be obtained by minimizing the expected error. This would be the solution of the following optimization problem:

$$\min \left\{ \sum_{i=1}^{k} \mathbb{E}\big[Z^i(\pi)\big] : \; Z^i(\pi) = \operatorname{dist}(\varphi(X^i;\pi), K_i), \; i = 1 \ldots k, \; \pi \in \mathcal{D}. \right\} \tag{3}$$

Here, a closed convex set $\mathcal{D} \subseteq \mathcal{R}^s$ describes the set of feasible parameters π. Our goal is to introduce a family of risk-averse classifiers, where the expectation is replaced by law-invariant coherent measures of risk.

We start with the formulation of an optimization problem for binary classification, in which the (estimated) expected total error is minimized.

$$\begin{aligned} \min_{v,\gamma,Z^1,Z^2} \; & \frac{1}{m_1}\sum_{j=1}^{m_1} z_j^1 + \frac{1}{m_2}\sum_{j=1}^{m_2} z_j^2 \\ \text{s. t.} \; & \langle v, x_j^1\rangle - \gamma + z_j^1 \geq 0, \quad j = 1, \ldots, m_1, \\ & \langle v, x_j^2\rangle - \gamma - z_j^2 \leq 0, \quad j = 1, \ldots, m_2, \\ & \|v\| = 1, \quad Z^1 \geq 0, \quad Z^2 \geq 0. \end{aligned} \tag{4}$$

In this formulation, Z^1 and Z^2 are random variables expressing the classification error for class 1 and class 2, respectively. Those variables have realizations z_j^1 and z_j^2. Note that z_i^1 and z_i^2 will satisfy Eq. (2) only if we use the Euclidean norm of v in (4).

The soft-margin SVM with parameters $M > 0$ and $\delta > 0$ is formulated as follows:

$$\begin{aligned} \min_{v,\gamma,Z^1,Z^2} \; & M\left(\sum_{j=1}^{m_1} z_j^1 + \sum_{j=1}^{m_2} z_j^2\right) + \delta\|v\|^2 \\ \text{s. t.} \; & \langle v, x_j^1\rangle - \gamma + z_j^1 \geq 1, \quad j = 1, \ldots, m_1, \\ & \langle v, x_j^2\rangle - \gamma - z_j^2 \leq -1, \quad j = 1, \ldots, m_2, \\ & Z^1 \geq 0, \quad Z^2 \geq 0. \end{aligned} \tag{5}$$

In problem (5), the normal vector v of the separating hyperplane can be of any positive length. Observe that multiplying the solution of problem (5), v and γ, by a positive constant does not change the separating plane. In problem (5), the estimated expected total classification error equals

$$\frac{1}{m_1\|v\|}\sum_{i=1}^{m_1} \max(z_i^1 - 1, 0) + \frac{1}{m_2\|v\|}\sum_{j=1}^{m_2} \max(z_j^2 - 1, 0)$$

This means that the objective function does not necessarily minimize the expected classification error although the variables z_j^1 and z_j^2 are indicative of misclassification occurrence.

We propose a new family of risk functionals: coherent measures of risk representing the point of view that *it should be possible to treat misclassification errors for each classes with different attitude to risk*. While the total expected error is a sum of expected misclassification in each class, the risk in a system measured by a coherent risk measure is not a sum of the risk of each component. That is why, we do not simply minimize the sum of risks for each class. We adopt a point of view on optimality of risk allocation as the one in risk sharing theory in mathematical finance. However, we emphasize that the problem setting and the results associated with risk sharing of losses in financial institutions are inapplicable to the classification problem as it will become clear in due course.

3 Coherent measures of risk

Measures of risk are widely used in finance and insurance. Additionally, the signal to noise measures, used in engineering and statistics (Fano factor Fano 1947 or the index of dispersion Cox and Lewis 1966) are of similar spirit. An axiomatic theory of measures of risk is presented in Ogryczak and Ruszczyński (1999), Artzner et al. (1999), Föllmer and Schied (2011), Kijima and Ohnishi (1993), Rockafellar et al. (2006). In a more general setting, risk measures are analyzed in Ruszczynski and Shapiro (2006). For $p \in [1, \infty]$ and a probability space $(\Omega, \mathcal{F}, P)$, we use the notation $\mathcal{L}_p(\Omega, \mathcal{F}, P)$, for the space of random variables with finite p-th moments.

Definition 2 A *coherent measure of risk* is a functional $\varrho : \mathcal{L}_p(\Omega) \to \mathbb{R}$ satisfying the following axioms:

Convexity: For all $X, Y, \gamma \in [0, 1]$, $\varrho(\gamma X + (1-\gamma)Y) \le \gamma\varrho(X) + (1-\gamma)\varrho(Y)$.
Monotonicity: If $X_\omega \ge Y_\omega$ for P-a.a $\omega \in \Omega$, then $\varrho(X) \ge \varrho(Y)$.
Translation Equivariance: For any $a \in \mathbb{R}$, $\varrho(X + a) = \varrho(X) + a$ for all X.
Positive Homogeneity: If $t > 0$ then $\varrho(tX) = t\varrho(X)$ for any X.

For an overview of the theory of coherent measures of risk, we refer to Shapiro et al. (2014) and the references therein.

A risk measure $\varrho(\cdot)$ is called *law-invariant* if $\varrho(X) = \varrho(Y)$ whenever the random variables X and Y have the same distributions. It is clear that in our context, only law invariant measures of risk are relevant.

The following result is known as a dual representation of coherent measures of risk (cf. Shapiro et al. 2014). The space $\mathcal{L}_p(\Omega)$ and the space $\mathcal{L}_q(\Omega)$ with $\frac{1}{p} + \frac{1}{q} = 1$ are viewed as paired vector spaces with respect to the bilinear form

$$\langle \zeta, Z \rangle = \int_\Omega \zeta(\omega) Z(\omega) dP(\omega), \quad \zeta \in \mathcal{L}_q(\Omega), \quad Z \in \mathcal{L}_p(\Omega). \tag{6}$$

For any $\zeta \in \mathcal{L}_p(\Omega)$, we can view $\langle \zeta, Z \rangle$ as the expectation $\mathbb{E}_Q[Z]$ taken with respect to the probability measure $dQ = \zeta dP$, defined by the density ζ, i.e., Q is absolutely continuous with respect to P and its Radon-Nikodym derivative is $dQ/dP = \zeta$. For any finite-valued coherent measure of risk ϱ, a convex subset $\mathcal{A}$ of probability density functions $\zeta \in \mathcal{L}_q(\Omega)$ exists, such that for any random variable $Z \in \mathcal{L}_p(\Omega)$, it holds

$$\varrho(Z) = \sup_{\zeta \in \mathcal{A}} \langle \zeta, Z \rangle = \sup_{dQ/dP \in \mathcal{A}} \mathbb{E}_Q[Z]. \tag{7}$$

This result reveals how measures of risk provide robustness with respect to the changes of the distribution. Their application constitutes a new approach to robust statistical inference.

For a random variable $X \in \mathcal{L}_p(\Omega)$ with distribution function $F_X(\eta) = P\{X \le \eta\}$, we consider its survival function $\bar{F}_X(\eta) = P(X > \eta)$ and the left-continuous inverse of the cumulative distribution function defined as follows:

$$F_X^{(-1)}(\alpha) = \inf\{\eta : F_X(\eta) \ge \alpha\} \quad \text{for} \quad 0 < \alpha < 1,$$

i.e., $F_X^{(-1)}(\alpha)$ is the left α-quantile of X.

We intend to investigate the distribution of classification errors and that is why we have a preference to small outcomes (small errors). Following Shapiro et al. (2014), the *Value at Risk* at level α of a random error X is defined by setting

$$\mathrm{VaR}_\alpha(X) = F_X^{(-1)}(1-\alpha).$$

The risk here is understood as the probability of the error X being large. This point of view corresponds to minimizing the probability of misclassification. Although Value at Risk is an intuitively appealing measure, it is not coherent.

In the theory of measures of risk, a special role is played by the functional called the Conditional Value-at-Risk and denoted $\mathrm{CVaR}(\cdot)$ (also known as Conditional Value at Risk or CVaR, see Ogryczak and Ruszczyński 2002; Rockafellar and Uryasev 2002). The *Conditional Value at Risk* of X at level α is defined as

$$\mathrm{CVaR}_\alpha(X) = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_t(X)\,dt. \tag{8}$$

For the second equality, we refer to Dentcheva and Martinez (2012), Ogryczak and Ruszczyński (2002). This is the representation (cf. also Shapiro et al. 2014) suitable for optimization problems. Due to Kusuoka theorem (Kusuoka 2001; Shapiro et al. 2014, Thm. 6.24), every law invariant, finite-valued coherent measure of risk on $\mathcal{L}^p(\Omega)$ for non-atomic probability space can be represented as a mixture of Conditional Value-at-Risk at all probability levels. This result can be extended for finite probability spaces with equally likely observations.

A popular class of coherent measures of risk include the semideviation of a random variable. The *upper* semideviation of order p is defined as

$$\sigma_p^+[Z] := \left(\mathbb{E}\left[\left(Z-\mathbb{E}[Z]\right)_+^p\right]\right)^{1/p}, \tag{9}$$

where $p \in [1,\infty)$ is a fixed parameter. It is well defined for all random variables Z with finite p-th order moments. The mean–upper-semideviation measure has the general form

$$\mathbb{E}[Z] + c\sigma_p^+[Z], \quad \text{for some constant } c \in [0,1]. \tag{10}$$

The mean-upper-semideviation measure reflects preferences to small realizations of the random variables and it aims at penalization of the excess over the expected value when Z depends on the choice of a decision maker and the choice is taken as to minimize this risk measure. This measure of risk is suited for our purposes since in the setting of misclassification small values are preferred. Higher value of the constant c corresponds to higher risk-aversion while $c = 0$ corresponds to the risk-neutral attitude.

In order to illustrate the connection to robust statistics and robust optimization, we provide the dual representation of the risk measures which we shall use in our numerical study. For parameters α, β, c, all contained in $[0,1]$ and $p \geq 1$, we have:

$$\begin{aligned}
&\mathrm{CVaR}_\alpha(X) = \sup\{\mathbb{E}_\zeta[Z] \,:\, \zeta \in \mathcal{L}_\infty(\Omega) : \zeta(\omega) \in [0,\alpha^{-1}] \text{ a.e.}\};\\
&(1-\beta)\mathbb{E}[Z] + \beta\mathrm{CVaR}_\alpha(Z)\\
&\quad = \sup\{\mathbb{E}_\zeta[Z] \,:\, \zeta \in \mathcal{L}_\infty(\Omega) : \zeta(\omega) \in [1-\beta, 1+\beta(1-\alpha)\alpha^{-1}] \text{ a.e.}\};\\
&\mathbb{E}[Z] + c\sigma_p^+[Z] = \sup\{\mathbb{E}_\zeta[Z] \,:\, \zeta \in \mathcal{L}_q(\Omega) : \zeta = 1+\xi-\mathbb{E}[\xi],\ \|\xi\|_q \leq c\}
\end{aligned} \tag{11}$$

Note that using these measures results in taking the worst expected misclassification when it is evaluated not only by the empirical probability mass function but all mass functions satisfying the conditions in (11).

Statistical estimators of spectral law-invariant measures of risk using Kusuoka representations are proposed in Dentcheva and Penev (2010). Furthermore, central limit theorems for general composite risk functionals, which incorporate the risk measures used in this paper are established in Dentcheva et al. (2017). Other classes of coherent measures of risk were pro-

posed and analyzed in Dentcheva et al. (2010), Krokhmal (2007), Ogryczak and Ruszczyński (2002), Shapiro et al. (2014) and the references therein.

4 Risk sharing in classification

First, we introduce the notion of a risk-averse classifier and show that risk-averse classifiers are associated with minimal points of the attainable errors, where the minimality is with respect to a suitable stochastic order. Let a set of labeled data, a parametric classifier family $\varphi(\cdot;\pi)$ with the associated collection of sets $K_i, i = 1\ldots,k$, and the law-invariant coherent risk measures $\varrho_i, i = 1\ldots,k$ be given. The presumption is that we have different attitude to misclassification risk in the various classes and the total risk is shared among the classes according to risk-averse preferences. Risk preferences in classification have been previously specified via cost-sensitive objective functions which may place different costs on the misclassification errors of each class. However, it is difficult to determine a proper weighting that truly reflects the risk preferences of the decision maker. In this section, we show that using risk functionals leads to cost-sensitive formulation, where the weighting is implied by the choice of risk measure.

First, we show that $Z^i(\pi) \in \mathcal{L}_p(\Omega,\mathcal{F},P)\ i = 1,\ldots,k$ under appropriate conditions.

Theorem 1 *Assume that $X^i \in \mathcal{L}_p(\Omega,\mathcal{F},P)$, K_i are closed convex sets, $i = 1,\ldots,k$, and that the function $\varphi(\cdot;\pi)$ satisfies the following growth condition*

$$\|\varphi(x,\pi)\| \le C_1(\pi) + C_2(\pi)\|x\|^p, \quad x \in \cup_{i=1}^k \operatorname{supp} X^i, \tag{12}$$

where $C_1(\pi)$ and $C_2(\pi)$ are constants depending on π and $\|\varphi(x,\pi)\|$ refers to the Euclidean norm in $\mathbb{R}^d$. Then $Z^i(\pi) \in \mathcal{L}_1(\Omega,\mathcal{F},P)$, $i = 1,\ldots,k$.

Proof The distance functions $z \mapsto \operatorname{dist}(z,K_i), i = 1,\ldots,k$, are continuous convex functions (see, e.g., Beer 1993) and $\operatorname{dist}(z,K_i) < \infty$ for all $z \in \mathbb{R}^n$. Furthermore, let $\hat{x}_i$ be the norm-minimal element of K_i. Then

$$\operatorname{dist}(z,K_i) \le \|z - \hat{x}_i\| \le \|z\| + \|\hat{x}_i\|.$$

For all $i = 1,\ldots,k$, we obtain the estimate over the support of the random vector X^i:

$$\operatorname{dist}(\varphi(x;\pi),K_i) \le \|\varphi(x;\pi)\| + \|\hat{x}_i\| \le C_1(\pi) + C_2(\pi)\|x\|^p + \|\hat{x}_i\|.$$

We conclude that $Z_i, i = 1,\ldots,k$, are integrable because $\|X^i\|_p$ is finite. □

Let $\mathcal{Y}$ denote the set of all random vectors $(Z^1(\pi),\ldots,Z^k(\pi))$ obtained as $Z^i(\pi) = \operatorname{dist}(\varphi(X^i;\pi),K_i)$ for some $\pi \in \mathcal{D}$, i.e., $\mathcal{Y}$ is the set of *all attainable classification errors* considered as random vectors in the corresponding probability space. In the classification problem, we deal with their representation from the available sample calculated as follows:

$$z^i_j(\pi) = \operatorname{dist}(\varphi(x_j;\pi),K_i), x_j \in S_i, \quad j = 1,\ldots,m_i, \quad i = 1,\ldots k.$$

for a given parameter $\pi \in \mathcal{D}$.

Definition 3 A vector $w \in \mathbb{R}^k$ represents an *attainable risk allocation* for the classification problem, if a parameter $\pi \in \mathcal{D}$ exists such that

$$w = \big(\varrho_1(Z^1(\pi)),\ldots,\varrho_k(Z^k(\pi))\big) \in \mathbb{R}^k \quad \text{for} \quad \big(Z^1(\pi),\ldots,Z^k(\pi)\big) \in \mathcal{Y}.$$

We denote the set of all attainable risk allocations by $\mathcal{X}$. Assume that a partial order on $\mathbb{R}^k$ is induced by a pointed convex cone $\mathcal{K} \subset \mathbb{R}^k$, i.e.,

$$v \preceq_{\mathcal{K}} w \text{ if and only if } w - v \in \mathcal{K}.$$

Recall that a point $v \in A \subset \mathbb{R}^k$ is called *$\mathcal{K}$-minimal point of the set A* if no point $w \in A$ exists such that $v - w \in \mathcal{K}$. If $\mathcal{K} = \mathbb{R}^k_+$, then the notion of $\mathcal{K}$-minimal points of a set corresponds to the well-known notion of Pareto-efficiency or Pareto-optimality in $\mathbb{R}^k$.

Definition 4 A classifier $\varphi(\cdot;\pi)$ is called $\mathcal{K}$-optimal risk-averse classifier, if its risk-allocation is a $\mathcal{K}$-minimal element of $\mathcal{X}$. If $\mathcal{K} = \mathbb{R}^k_+$, then the classifier is called Pareto-optimal.

From now on, we focus on Pareto-optimality, but our results are extendable to the case of more general orders defined by pointed cones.

Definition 5 A *risk-sharing classification problem* (RSCP) is given by the set of labeled data, a parametric classifier family $\varphi(\cdot;\pi)$ with the associated collection of sets K_i, $i = 1\ldots,k$, and a set of law-invariant risk measures ϱ_i, $i = 1\ldots,k$. The risk-sharing classification problem consists of identifying a parameter $\pi \in \mathcal{D}$ resulting in a Pareto-optimal classifier $\varphi(\cdot;\pi)$.

We shall see that the Pareto-minimal risk allocations are produced by random vectors, which are minimal points in the set $\mathcal{Y}$ with respect to the usual stochastic order, defined next.

Definition 6 A random variable Z is stochastically larger than a random variable Z' with respect to the usual stochastic order (denoted $Z \succeq_{(1)} Z'$), if

$$\mathbb{P}(Z > \eta) \geq \mathbb{P}(Z' > \eta) \quad \forall\, \eta \in \mathbb{R}, \tag{13}$$

or, equivalently, $F_Z(\eta) \leq F_{Z'}(\eta)$. The relation is strict (denoted $Z \succ_{(1)} Z'$), if additionally, inequality (13) is strict for some $\eta \in \mathbb{R}$.

A random vector $\mathbf{Z} = (Z_1, \ldots Z_k)$ is stochastically larger than a random vector $\mathbf{Z}' = (Z'_1, \ldots Z'_k)$ (denoted $\mathbf{Z} \succeq \mathbf{Z}'$) if $Z_i \succeq_{(1)} Z'_i$ for all $i = 1, \ldots k$. The relation is strict if for some component $Z_i \succ_{(1)} Z'_i$.

The random vectors of $\mathcal{Y}$, which are non-dominated with respect to this order will be called *minimal points of* $\mathcal{Y}$.

For more information on stochastic orders see, e.g., Shaked and Shanthikumar (2007).

The following result is known for atomless probability spaces. We verify it for a sample space in order to deal with the empirical distributions.

Theorem 2 *Suppose the probability space $(\Omega, \mathcal{F}, P)$ is finite with equal probabilities of all simple events. Then every law-invariant risk functional ϱ is consistent with the usual stochastic order if and only if it satisfies the monotonicity axiom. If ϱ is strictly monotonic with respect to the almost sure relation, then ϱ is consistent with the strict dominance relation, i.e. $\varrho(Z_1) < \varrho(Z_2)$ whenever $Z_2 \succ_{(1)} Z_1$.*

Proof Assuming that $\Omega = \{\omega_1, \ldots, \omega_m\}$, let the random variable $U(\omega_i) = \frac{i}{m}$ for all $i = 1, \ldots, m$. If $Z_2 \succeq_{(1)} Z_1$, then defining $\hat{Z}_1 := F^{-1}_{Z_1}(U)$ and $\hat{Z}_2 := F^{-1}_{Z_2}(U)$, we obtain $\hat{Z}_2(\omega) \geq \hat{Z}_1(\omega)$ for all $\omega \in \Omega$. Due to the monotonicity axiom, $\varrho(\hat{Z}_2) \geq \varrho(\hat{Z}_1)$. The random variables $\hat{Z}_i$ and Z_i, $i = 1, 2$, have the same distribution by construction. This entails that $\varrho(Z_2) \geq \varrho(Z_1)$ because the risk measure is law invariant. Consequently, the risk measure ϱ is consistent with the usual stochastic order. The other direction is straightforward. □

This observation justifies our restriction to risk measures, which are consistent with the usual stochastic order, also known as the first order stochastic dominance relation. Furthermore, when dealing with non-negative random variables as in the context of classification, then strictly monotonic risk measures associate no risk only when no misclassification occurs, as shown by the following statement.

Lemma 1 *If ϱ is a law invariant strictly monotonic coherent measure of risk, then*

$$\begin{aligned} &\varrho(Z) > 0 \quad \text{for all random variables } Z \geq 0 \ a.s., \quad Z \not\equiv 0 \\ &\varrho(Z) < 0 \quad \text{for all random variables } Z \leq 0 \ a.s., \quad Z \not\equiv 0. \end{aligned} \tag{14}$$

Proof Denote the random variable, which is identically equal zero by $\mathbf{0}$. Notice that $\varrho(\mathbf{0}) = \varrho(2\cdot\mathbf{0}) = 2\varrho(\mathbf{0})$, which implies that $\varrho(\mathbf{0}) = 0$. If $Z \geq 0$ a.s. and $Z \not\equiv 0$, then $\varrho(Z) > \varrho(\mathbf{0}) = 0$ by the strict monotonicity of ϱ. The second statement follows analogously. □

This statement implies that $\varrho_i(Z^i(\pi)) \geq 0$, $i = 1, \dots k$, for all $\pi \in \mathcal{D}$ and, therefore, the attainable allocations lie in the positive orthant, i.e., $\mathcal{X} \subseteq \mathbb{R}^k_+$. Consequently, the set $\mathcal{X}$ has minimal elements with respect to the Pareto-order. From now on, we adopt the following assumptions:

(A1) The risk measures ϱ^i used for evaluation of classification errors in classes $i = 1, ..., k$ are coherent, law invariant, and finite-valued.
(A2) The sets K_i, $i = 1, \dots k$ and $\mathcal{D} \subseteq \mathbb{R}^s$ are non-empty, closed and convex.
(A3) The function $\varphi(\cdot; \pi)$ satisfies the growth condition (12).

We point out that Assumptions (A2) and (A3) are satisfied for the examples, given in Sect. 1.

Theorem 3 *Assume (A1)–(A3). If the function $\varphi(x, \cdot)$ is continuous for every argument $x \in \mathbb{R}^n$, then the components of the attainable risk allocations $\varrho_i(Z^i(\cdot))$, $i = 1, \dots k$, are continuous functions. If additionally, each component of the vector function $\varphi(x, \cdot)$ is an affine function, then $\varrho_i(Z^i(\cdot))$, $i = 1, \dots k$ are convex functions.*

Proof Recall again that the distance functions $z \mapsto \operatorname{dist}(z, K_i)$ are continuous convex functions and $\operatorname{dist}(z, K_i) < \infty$ for all $z \in \mathbb{R}^n$. Thus, the composition of the distance function with the continuous function $\varphi(x; \cdot)$ is continuous, meaning that the random variable $Z^i(\pi) = \operatorname{dist}(\varphi(X^i; \pi), K_i)$ has realizations, which are continuous functions of π. The variables Z^i are integrable due to Theorem 1. Therefore, $Z^i(\cdot)$ is continuous with respect to the norm in the space $\mathcal{L}_1(\Omega)$. Since the risk measures $\varrho_i(\cdot)$ are convex and finite, they are continuous on $\mathcal{L}_1$. We conclude that its composition with the risk measure: $\varrho_i(Z^i(\cdot))$, is continuous.

In order to prove convexity, let $\lambda \in (0, 1)$ and let $\pi_\lambda = \lambda\pi + (1 - \lambda)\pi'$.

Let $z^i(\pi), z^i(\pi') \in K_i$ be the points such that

$$\|\varphi(x; \pi) - z^i(\pi)\| = \min_{z \in K_i} \|\varphi(x; \pi) - z\| \tag{15}$$

$$\|\varphi(x; \pi) - z^i(\pi')\| = \min_{z \in K_i} \|\varphi(x; \pi') - z\| \tag{16}$$

We define $z_\lambda = \lambda z^i(\pi) + (1 - \lambda) z^i(\pi')$. Due to the convexity of K_i, we have $z_\lambda \in K_i$. As $\varphi(x, \cdot)$ is affine, we obtain

$$\varphi(x; \pi_\lambda) = \lambda\varphi(x; \pi) + (1 - \lambda)\varphi(x; \pi').$$

This entails the following inequality for all $i = 1, \dots k$ and all $z \in \mathbb{R}^d$:

$$\begin{aligned}\min_{z\in K_i} \|\varphi(x;\pi_\lambda) - z\| \le \|\varphi(x;\pi_\lambda) - z^i_\lambda\| &= \|\varphi(x;\pi_\lambda) - \lambda z^i(\pi) - (1-\lambda)z^i(\pi')\| \\ &= \|\lambda\big(\varphi(x;\pi) - z^i(\pi)\big) + (1-\lambda)\big(\varphi(x;\pi') - z^i(\pi')\big)\| \\ &\le \lambda\|\varphi(x;\pi) - z^i(\pi)\| + (1-\lambda)\|\varphi(x;\pi') - z^i(\pi'))\| \\ &= \lambda \min_{z\in K_i} \|\varphi(x;\pi) - z\| + (1-\lambda) \min_{z\in K_i}(\varphi(x;\pi') - z)\|.\end{aligned}$$

Therefore,

$$\mathrm{dist}(\varphi(x;\pi_\lambda), K_i) \le \lambda \mathrm{dist}(\varphi(x;\pi), K_i) + (1-\lambda)\mathrm{dist}(\varphi(x;\pi'), K_i).$$

The monotonicity and convexity axioms for the risk measures imply that

$$\begin{aligned}&\varrho_i\big(\mathrm{dist}(\varphi(X;\pi_\lambda), K_i)\big) \\ &\quad\le \lambda\varrho_i\big(\mathrm{dist}(\varphi(X;\pi), K_i)\big) + (1-\lambda)\varrho_i\big(\mathrm{dist}(\varphi(X;\pi'), K_i)\big)\end{aligned}$$

for all $i = 1, \dots, k$. □

This result implies the existence of Pareto-optimal classifier. Furthermore, the convexity property allows us to identify the Pareto-optimal risk-allocations by using scalarization techniques.

Corollary 1 *Assume (A1)–(A3) and let the function $\varphi(x, \cdot)$ be affine for every argument $x \in \mathbb{R}^n$. Then a parameter π defines a Pareto-optimal classifier $\varphi(\cdot, \pi)$ for the given RSCP if and only if a scalarization vector $w \in \mathbb{R}^k_+$ exists with $\sum_{i=1}^k w_i = 1$, such that π is a solution of the problem*

$$\min_{\pi\in\mathcal{D}} \sum_{i=1}^k w_i \varrho_i\big(\mathrm{dist}(\varphi(X_i;\pi), K_i)\big). \tag{17}$$

Proof Statement follows form the well-known scalarization theorem in vector optimization problems (Miettinen 1999) and Theorem 3. □

Theorem 4 *Assume that the risk measures ϱ_i are law invariant and strictly monotonic for all $i = 1, \dots k$. If a classifier $\varphi(\cdot;\pi)$ is Pareto-optimal, then its corresponding random vector $(Z^1(\pi), \dots, Z^k(\pi))$ is a minimal point of $\mathcal{Y}$ with respect to the order of Definition 6.*

Proof Suppose that $\varphi(\cdot;\pi)$ is Pareto-optimal and the point $Z(\pi) = (Z^1(\pi), \dots, Z^k(\pi))$ is not minimal. Then a parameter π' exists, such that the corresponding vector $Z(\pi')$ is strictly stochastically dominated by Z, which implies $Z^i(\pi) \succeq_{(1)} Z^i(\pi')$ with a strict relation for some component. We obtain $\varrho_i(Z^i(\pi)) \ge \varrho_i(Z^i(\pi'))$ for all $i = 1, \dots, k$ with a strict inequality for some i due to the consistency of the coherent measures of risk with the strong stochastic order relation, which contradicts the Pareto-optimality of $\varphi(\cdot;\pi)$. □

We consider the sample space $\Omega = \prod_{i=1}^k \Omega_i$ where $(\Omega_i, \mathcal{F}_i, P_i)$ is a finite space with m_i simple events $\omega_j \in \Omega_i$, $P_i(\omega_j) = \frac{1}{m_i}$, and $\mathcal{F}_i$ consisting of all subsets of Ω_i.

Theorem 5 *Assume (A1)–(A3). Suppose each component of the vector function $\varphi(x, \cdot)$ is affine for every $x \in \mathbb{R}^n$. If the parameter $\hat{\pi}$ defines a Pareto-optimal classifier $\varphi(\cdot, \hat{\pi})$ for*

the RSCP, then a probability measure μ on Ω exists so that $\hat{\pi}$ is an optimal solution for the problem

$$\min_{\pi \in \mathcal{D}} \sum_{i=1}^{k} \sum_{j=1}^{m_i} \mu_j^i \operatorname{dist}(\varphi(x_j^i; \pi), K_i). \tag{18}$$

Proof Since the parameter $\hat{\pi}$ defines a Pareto-optimal classifier $\varphi(\cdot, \hat{\pi})$ for the RSCP and all conditions of Corollary 1 are satisfied, then $\hat{\pi}$ is an optimal solution of problem (17) for some scalarization w. Let $\mathcal{A}_i$ denotes the set of probability measures corresponding to the risk measure ϱ_i, $i = 1, \ldots, k$ in representation (7). Since the risk measures ϱ_i take finite values on Ω_i, the sets $\mathcal{A}_i$ are non-empty and compact. Thus, the supremum in the dual representation (7) is achieved at some elements $\zeta^i \in \mathcal{A}_i$. We have $\zeta_j^i \geq 0$, $\sum_{j=1}^{m_i} \frac{\zeta_j^i}{m_i} = 1$ because ζ_i are probability densities. We obtain

$$\varrho_i(\operatorname{dist}(\varphi(X^i; \pi), K_i)) = \sum_{j=1}^{m_i} \frac{\zeta_j^i}{m_i} \operatorname{dist}(\varphi(x_j^i; \pi), K_i).$$

Setting $\mu_j^i = w_i \frac{\zeta_j^i}{m_i}$, $j = 1, \ldots, m_i$, $i = 1, \ldots, k$, we observe that the vector $\mu \in R^{m_1 + \cdots m_k}$ constitutes a probability mass function. Thus, problem (17) can be reformulated as (18). □

This result shows that the RSCP can be viewed as a classification problem in which the expectation error is minimized. However, the expectation is not calculated with respect to the empirical distribution but with respect to another measure μ, which is implicitly determined by the chosen measures of risk. It is the worst expectation according to our risk-averse preferences, which are represented by the choice of the measures ϱ_i, $i = 1, \ldots, k$.

The composite nature of the problem (17) is difficult and that is why we reformulate the problem. We introduce auxiliary variables $Y \in \mathcal{L}_1(\Omega, \mathcal{F}, P; \mathbb{R}^m)$, $i = 1, \ldots k$, which are defined by the constraints:

$$\varphi(X^i; \pi) + Y^i \in K_i \quad \forall i = 1, \ldots, k.$$

Problem (17) can be reformulated to

$$\begin{aligned} \min_{\pi, Y} \ & \sum_{i=1}^{k} w_i \varrho_i(\|Y^i\|) \\ \text{s.t. } & \varphi(X^i; \pi) + Y^i \in K_i, \quad \forall i = 1, \ldots, k, \\ & \pi \in \mathcal{D}. \end{aligned} \tag{19}$$

We show that this problem is equivalent to (17).

Lemma 2 *Assume that K_i, $i = 1, \ldots, k$, are non-empty, closed convex sets. For any solution $\hat{\pi}$ of problem (17), random vectors $\hat{Y}^i$ exist, so that $(\hat{\pi}, \hat{Y})$ solves problem (19) as well, where $\hat{Y} = (\hat{Y}^k, \ldots, \hat{Y}^k)$ and for any solution $(\hat{\pi}, \hat{Y})$ of problem (19), the vector $\hat{\pi}$ is a solution of problem (17) as well.*

Proof Observe that for any fixed point $\pi \in \mathcal{D}$, the function $\sum_{i=1}^{k} w_i \varrho_i(\|Y^i\|)$ achieves minimal value with respect to the constraints on the variables Y^i using the projections of the realizations of X^i onto K_i:

$$Y^i(\omega) = \mathrm{Proj}_{K_i}\big((\varphi(X(\omega);\pi)\big) - \varphi(X(\omega);\pi). \tag{20}$$

Here $\mathrm{Proj}_{K_i}(z)$ denotes the Euclidean projection of the point z onto the set K_i. Then, $\|Y^i\| = \mathrm{dist}(\varphi(X^i;\pi), K_i)$ and the objective functions of both problems have the same value. Therefore, the minimal value is achieved at the same point $\hat{\pi}$ and the corresponding $\hat{Y}^i_j$ is obtained from Eq. (20). □

Recall that the normal cone to a set $\mathcal{D} \subset \mathbb{R}^s$ is defined as

$$\mathcal{N}_{\mathcal{D}}(\pi) = \{a \in \mathbb{R}^s : \langle a, d - \pi\rangle \le 0 \ \text{ for all } d \in \mathcal{D}\}.$$

For brevity, we denote the normal cone to the feasible set of problem (19) by $\mathcal{N}$ and the normal cones to the sets K_i by $\mathcal{N}_i$, $i = 1,\dots,k$. We formulate optimality conditions for problem (19).

We denote the realizations of the random vectors Y^i, $i = 1,\dots,k$ when π is used, by $y^i_j(\pi)$, $j = 1,\dots m_i$, $i = 1,\dots,k$. More precisely, we have

$$y^i_j(\pi) = \mathrm{Proj}_{K_i}\big((\varphi(x^i_j;\pi)\big) - \varphi(x^i_j;\pi) \quad j = 1,\dots m_i,\ i = 1,\dots,k.$$

We suppress the argument π whenever it does not lead to confusion. Additionally, we denote the Jacobian of φ with respect to π by $D\varphi(x;\pi)$. Consider the sample-based version of problem (19):

$$\begin{aligned} \min_{\pi,Y}\ & \sum_{i=1}^{k} w_i \varrho_i(\|Y^i\|) \\ \text{s.t. }\ & \varphi(x^i_j;\pi) + y^i_j \in K_i, \quad \forall j = 1,\dots,m_i,\ i = 1,\dots,k, \\ & \pi \in \mathcal{D}. \end{aligned} \tag{21}$$

Theorem 6 *Assume that the sets K_i, $i = 1,\dots,k$ are closed convex polyhedral cones and $\varphi(x;\cdot)$ is an affine vector function. A feasible point $(\hat{\pi}, \hat{Y})$ is optimal for problem (21) if and only if probability mass functions $\zeta^i \in \partial\varrho_i(0)$ and vectors g^i_j from $\partial\|\hat{y}^i_j\|$ exist such that*

$$0 \in -\sum_{i=1}^{k}\sum_{j=1}^{m_i} w_i \zeta^i_j (g^i_j)^\top D\varphi(X^i;\hat{\pi}) + \mathcal{N}_{\mathcal{D}}(\hat{\pi}) \tag{22}$$

$$w_i \zeta^i_j g^i_j \in \mathcal{N}_i\big(\varphi(x^i_j;\hat{\pi}) + \hat{y}^i_j\big) \quad \textit{for all} \ \ j = 1,\dots m_i, \quad i = 1,\dots k. \tag{23}$$

Proof We assign Lagrange multipliers λ^i_j to the inclusion constraints and define the Lagrange function as follows:

$$L(\pi, Y, \lambda) = \sum_{i=1}^{k}\left(w_i \varrho_i(\|Y^i\|) + \sum_{j=1}^{m_i} \big\langle \varphi(x^i_j;\pi) + y^i_j, \lambda^i_j \big\rangle \right).$$

Using optimality conditions, we obtain that $(\hat{\pi}, \hat{Y})$ is optimal for problem (21) if and only if $\hat{\lambda}$ exists such that

$$0 \in \partial_{(\pi,Y)} L(\hat{\pi}, \hat{Y}, \hat{\lambda}) + \mathcal{N}(\hat{\pi}, \hat{Y})$$
$$\hat{\lambda}_j^i \in \mathcal{N}_i\big(\varphi(x_j^i; \hat{\pi}) + \hat{y}_j^i\big).$$

Considering the partial derivatives of the Lagrangian with respect to the two components, we obtain

$$0 \in \sum_{i=1}^{k} \sum_{j=1}^{m_i} (\hat{\lambda}_j^i)^\top D\varphi(x_j^i; \hat{\pi}) + \mathcal{N}_{\mathcal{D}}(\hat{\pi}) \tag{24}$$

$$0 = w_i \partial_Y \varrho_i(\|Y\|) + \hat{\lambda}^i, \quad i = 1, \dots k, \tag{25}$$

$$\hat{\lambda}_j^i \in \mathcal{N}_i\big(\varphi(x_j^i; \hat{\pi}) + \hat{y}_j^i\big), \quad j = 1, \dots, m_i, \ i = 1, \dots k. \tag{26}$$

We calculate the multipliers $\hat{\lambda}^i$ from the Eq. (25) using elements $\zeta^i \in \partial \varrho_i(0)$ and g_j^i from $\partial \|\hat{y}_j^i\|$. We obtain:

$$\hat{\lambda}_j^i = -w_i \zeta_j^i g_j^i, \quad j = 1, \dots, m_i, \ i = 1, \dots k.$$

Notice that $g_j^i = \frac{\hat{y}_j^i}{\|\hat{y}_j^i\|}$ whenever $\hat{y}_j^i \neq 0$, otherwise $g_j^i \in \mathbb{R}^d$ can be any vector with $\|g_j^i\| \leq 1$. Substituting the value of $\hat{\lambda}^i$ into (24) and (26), we obtain condition (22) and (23). □

We note that, we can define again a probability mass function μ by setting $\mu_j^i = w_i \zeta_j^i$ and interpret the Karush–Kuhn–Tucker condition as follows:

$$\mathbb{E}_\mu (g_j^i)^\top D\varphi(X^i; \hat{\pi}) \in \mathcal{N}_{\mathcal{D}}(\hat{\pi})$$
$$\mu_j^i g_j^i \in \mathcal{N}_i\big(\varphi(x_j^i; \hat{\pi}) + \hat{y}_j^i\big) \text{ for all } \ j = 1, \dots m_i, \ i = 1, \dots k.$$

Problem (21) can be reformulated as a risk-averse two-stage optimization problem (cf. Shapiro et al. 2009). The first stage decision is π and the first stage problem is

$$\min_{\pi \in \mathcal{D}} \sum_{i=1}^{k} w_i \varrho_i\big(Z^i(\pi)\big)\big). \tag{27}$$

Given π, the calculation of each realization of $Z^i(\pi)$ amounts to solving the following problem

$$z_j^i(\pi) = \min_{y \in K_i} \|\varphi(x_j^i; \pi) - y\|, \quad j = 1, \dots m_i, \ i = 1, \dots k. \tag{28}$$

Calculating $z_j^i(\pi)$ might be very easy for specific regions K_i such as the cones in the example of the polyhedral classifier. Every component of the solution vector $\hat{z}_j^i$ to problem (28) can be computed as follows:

$$(\hat{z}_j^i)_\ell = \begin{cases} \max\{0, -(\varphi(x_j^i; \pi))_\ell\} & \text{for } \ell = i; \\ \max\{0, (\varphi(x_j^i; \pi))_\ell\} & \text{for } \ell \neq i; \end{cases} \quad \ell = 1, \dots, k. \tag{29}$$

Then the optimal value of (28) is

$$z_j^i(\pi) = \left(\sum_{\ell=1}^{k} (\hat{z}_j^i)_\ell^2 \right)^{\frac{1}{2}}.$$

This point of view facilitates the application of stochastic optimization methods to solve the problem.

5 Confidence intervals for the risk

In this section, we analyze the risk-averse classification problem when we increase the data sets and derive confidence intervals for the misclassification risk. We use the results on statistical inference for composite risk functionals presented in Dentcheva et al. (2017). In Dentcheva et al. (2017), a composite risk functional is defined in the following way.

$$\varrho(X) = \mathbb{E}\left[f_1\left(\mathbb{E}\left[f_2\left(\mathbb{E}\left[\cdots f_\ell\left(\mathbb{E}\left[f_{\ell+1}(X)\right], X\right)\right]\cdots, X\right)\right], X\right)\right] \tag{30}$$

where X is an n-dimensional random vector with unknown distribution, P_X. The functions f_j are such that $f_j(\eta_j, x) : \mathbb{R}^{n_j} \times \mathbb{R}^n \to \mathbb{R}^{n_{j-1}}$ for $j = 1, \ldots, \ell$ and $n_0 = 1$. The function $f_{\ell+1}$ is such that $f_{\ell+1}(x) : \mathbb{R}^n \to \mathbb{R}^{n_\ell}$.

A law-invariant risk-measure $\varrho(X)$ is an unknown characteristic of the distribution P_X. The empirical estimate of $\varrho(X)$ given N independent and identically distributed observations of X is given by the plug-in estimate

$$\varrho^{(N)} = \sum_{i_0=1}^{N} \frac{1}{N}\left[f_1\left(\sum_{i_1=1}^{N}\frac{1}{N}\left[f_2\left(\sum_{i_2=1}^{N}\frac{1}{N}\left[\cdots f_\ell\left(\sum_{i_\ell=1}^{N}\frac{1}{N}f_{\ell+1}(X_{i_\ell}), X_{i_{\ell-1}}\right)\right]\right.\right.\right.$$
$$\left.\left.\left.\cdots, X_{i_1}\right)\right], X_{i_0}\right)\right] \tag{31}$$

It is shown in Dentcheva et al. (2017) that the most popular measures of risk fit the structure (30). It is established that the plug-in estimator satisfies a central limit formula and the limiting distribution is described. This is the distribution of the Hadamard-directional derivative of the risk functional ϱ when a normal random variable is plugged in. Recall the notion of Hadamard directional derivatives of the functions $f_j(\cdot, x)$ at points μ_{j+1} in directions ζ_{j+1}. It is given by

$$f_j'(\mu_{j+1}, x; \zeta_{j+1}) = \lim_{\substack{t\downarrow 0 \\ s\to\zeta_{j+1}}} \frac{1}{t}\left[f_j(\mu_{j+1} + ts, x) - f_j(\mu_{j+1}, x)\right].$$

The central limit formula holds under the following conditions:

(i) $\int \|f_j(\eta_j, x)\|^2 \, P(dx) < \infty$ for all $\eta_j \in I_j$, and $\int \mathrm{dist}^2(\varphi(X^i; \pi), K_i) P(dx) < \infty$;
(ii) For all realizations x of X^i, the functions $f_j(\cdot, x)$, $j = 1, \ldots, \ell$, are Lipschitz continuous:

$$\|f_j(\eta_j', x) - f_j(\eta_j'', x)\| \le \gamma_j(x)\|\eta_j' - \eta_j''\|, \quad \forall\, \eta_j', \eta_j'',$$

and $\int \gamma_j^2(x) \, P(dx) < \infty$.
(iii) For all realizations x of X^i, the functions $f_j(\cdot, x)$, $j = 1, \ldots, \ell$, are Hadamard directionally differentiable.

These properties are satisfied for the mean-semideviation risk measures as shown in Dentcheva et al. (2017). Furthermore, it is shown that similar construction represents the Conditional-Value-at-Risk.

For every parameter π the risk of misclassification for a given class $i = 1, \ldots, k$ can be fit to the setting (30) by choosing the innermost function $f_{\ell+1}(x) : \mathbb{R}^d \to \mathbb{R}$ to be $f_{\ell+1}(x) = \operatorname{dist}(\varphi(x;\pi), K_i)$ whenever φ satisfies properties (i)–(iii).

In our setting, each misclassification risk $\varrho_i\Big(\operatorname{dist}\big(\varphi(X^i;\pi), K_i\big)\Big)$ is estimated by $\varrho_i^{(m_i)}\big(\|\hat{Y}^i\|\big)$, where $(\hat{Y}^i;\hat{\pi})$ is the solution of problem (21). Denoting the estimated variance of the limiting distribution of $\varrho_i^{(m_i)}\big(\|\hat{Y}^i\|\big)$ (briefly $\varrho_i^{(m_i)}$) by σ_i^2, we obtain the following confidence interval:

$$\left[\ \varrho_i^{(m_i)} - t_{\alpha,\mathrm{df}}\frac{\sigma_i}{\sqrt{m_i}},\quad \varrho_i^{(m_i)} + t_{\alpha,\mathrm{df}}\frac{\sigma_i}{\sqrt{m_i}}\ \right].$$

Here α is the desired level of confidence, $t_{\alpha,\mathrm{df}}$ is the corresponding quantile of the t-distribution with degrees of freedom df. The degrees of freedom depend on the choice if risk measure and can be calculated as $df = m_i - \ell$, where ℓ is the number of compositions in formula (31). The decrease of the degrees of freedom form m_i is due to the estimation of the expected value associated with each composition. The total risk is estimated by

$$\hat{\varrho} = \sum_{i=1}^{k} w_i \varrho_i^{(m_i)}\big(\|\hat{Y}^i\|\big).$$

We obtain that $\hat{\varrho}$ has an approximately normal distribution with expected value ϱ and variance $\sum_{i=1}^{k}\frac{w_i^2\sigma_i^2}{m_i}$. A confidence interval for the entire risk ϱ, associated with the optimal classifier, is given by

$$\left[\ \hat{\varrho} - t_{\alpha,\mathrm{df}}\sqrt{\sum_{i=1}^{k}\frac{w_i^2\sigma_i^2}{m_i}},\quad \hat{\varrho} + t_{\alpha,\mathrm{df}}\sqrt{\sum_{i=1}^{k}\frac{w_i^2\sigma_i^2}{m_i}}\ \right].$$

We can use the confidence interval to evaluate how well the risk is estimated. The quantity $\frac{w_i^2\sigma_i^2}{m_i}$ gives us guidance about the need of additional observations from class i, which would help to reduce effectively the size of confidence interval of the risk, thus, will help us to evaluate the risk more precisely.

6 Risk sharing in SVM

We analyze the SVM problem in more detail. We consider only strictly monotonic coherent measures of risk ϱ_1, ϱ_2 for the two classes S_1 and S_2.

The *risk-sharing SVM problem (RSSVM)* consists in identifying a parameter $\pi = (v, \gamma) \in \mathbb{R}^n$ corresponding to a Pareto-minimal point of the attainable risk-allocations for the affine classifier $\varphi(z;\pi) = \langle v, z\rangle - \gamma$. Due to Corollary 1, we can determine a risk-averse classifier by solving the following problem:

$$\min_{v,\gamma,Z^1,Z^2}\ \lambda\varrho_1(Z^1) + (1-\lambda)\varrho_2(Z^2) \tag{32}$$

$$\text{s. t. } \langle v, x_j^1\rangle - \gamma + z_j^1 \geq 0, \quad j = 1, \ldots, m_1, \tag{33}$$

$$\langle v, x_j^2\rangle - \gamma - z_j^2 \leq 0, \quad j = 1, \ldots, m_2, \tag{34}$$

$$\langle v, v\rangle = 1, \tag{35}$$

$$Z^1 \geq 0, \quad Z^2 \geq 0. \tag{36}$$

Here $\lambda \in (0, 1)$ is a parameter representing the scalarization and is indicative of the risk sharing between the classes. The random variables Z^i can be represented by a deterministic vectors stacking all realizations z^i_j as components of it. Abusing notation, we shall use Z^i also for those vectors in $\mathbb{R}^{m_i}$, $i = 1, 2$.

We note that the normalization of the vector v automatically bounds γ because for any fixed v, the component γ can be considered restricted in a compact set $[\gamma_m(v), \gamma_M(v)]$, where

$$\gamma_M = \max_{1\le j\le m_i,\ i=1,2} v^\top x^i_j \quad \gamma_m = \min_{1\le j\le m_i,\ i=1,2} v^\top x^i_j. \tag{37}$$

Thus, in this case, we can set $\mathcal{D} = \mathbb{R}^n$.

We also consider a soft-margin risk-averse SVM based on problem (4), although the classification error might not be calculated properly. The problem reads

$$\min_{v,\gamma,Z^1,Z^2} \left\{\lambda\varrho_1(Z^1) + (1-\lambda)\varrho_2(Z^2) + \delta\|v\|^2 \colon (33), (34), (36)\right\} \tag{38}$$

In this problem, $\delta > 0$ is a small number. The objective function grows to infinity when the norm of v increases. Thus, we do not need to bound the norm of the vector v. It also automatically bounds γ, similar to problem (32)–(36).

We obtain a counterpart of the result in Jouini et al. (2008) for the risk sharing of random losses among constituents. We observe that the parameter (v, γ) for each Pareto-optimal classifier can be obtained by solving the following problem:

$$\begin{aligned} \min_{v,\gamma,Z^1,Z^2}\ & \varrho_1(Z^1) + \varrho_2(Z^2) \\ \text{s. t. }\ & \langle v, x^1_i\rangle - \gamma + \frac{1}{\lambda} z^1_i \ge 0, \quad i = 1, \dots, m_1, \\ & \langle v, x^2_j\rangle - \gamma - \frac{1}{1-\lambda} z^2_j \le 0, \quad j = 1, \dots, m_2, \\ & \langle v, v\rangle = 1, \quad Z^1 \ge 0, \quad Z^2 \ge 0. \end{aligned} \tag{39}$$

Lemma 3 *Problem* (39) *is equivalent to problem* (32)–(36).

Proof The equivalence follows from the axiom of positive homogeneity for the risk measures:

$$\lambda\varrho_1(Z^1) = \varrho_1(\lambda Z^1) \quad \text{and} \quad (1-\lambda)\varrho_2(Z^2) = \varrho_2((1-\lambda)Z^2).$$

Defining new random variables $\tilde{Z}^1 = \lambda Z^1$ and $\tilde{Z}^2 = (1-\lambda)Z^2$, we can rescale the variables in their respective inequality constraint. □

Although problem (38) is non-convex due to the presence of constraint (35), we can solve it by a dedicated numerical method using sequentially local convex approximations. Let $\bar{v}$ be a fixed point. The non-convex constraint can be approximated locally by using Taylor expansion:

$$\langle v, v\rangle - 1 \approx \langle \bar{v}, \bar{v}\rangle - 1 + 2\langle \bar{v}, v - \bar{v}\rangle = 2\langle \bar{v}, v\rangle - \langle \bar{v}, \bar{v}\rangle - 1.$$

If $\|\bar{v}\| = 1$, then we obtain:

$$\langle v, v\rangle - 1 \approx 2(\langle \bar{v}, v\rangle - 1).$$

For the sake of brevity, we denote the objective function by f:

$$f(v, \gamma, Z^1, Z^2) = \lambda\varrho_1(Z^1) + (1-\lambda)\varrho_2(Z^2)$$

Observe that for a feasible vector (v, γ) with $\|v\| \neq 1$, the misclassification errors are $\frac{1}{\|v\|} Z^1$ and $\frac{1}{\|v\|} Z^2$. Thus, using the positive homogeneity property of the risk measures, the true risk of misclassification is

$$\frac{1}{\|v\|}\big(\lambda \varrho_1(Z^1) + (1-\lambda)\varrho_2(Z^2)\big),$$

We propose the following method for solving problem (38).

Risk averse binary classification method
Step 0 Set $\ell = 1$, $v^0 = \frac{1}{m_1}\sum_{i=1}^{m_1} x_i^1 - \frac{1}{m_2}\sum_{i=1}^{m_2} x_i^2$, calculate γ^0, $Z^{1,0}$, and $Z^{2,0}$ as a solution of the following problem

$$\begin{aligned} \min_{\gamma, Z^1, Z^2} \; & f(v^0, \gamma, Z^1, Z^2) \\ \text{s. t. } & \langle v^0, x_i^1\rangle - \gamma + z_i^1 \geq 0, \quad i = 1, \ldots, m_1, \\ & \langle v^0, x_j^2\rangle - \gamma - z_j^2 \leq 0, \quad j = 1, \ldots, m_2. \end{aligned} \tag{40}$$

Step 1 Solve the convex approximation problem

$$\begin{aligned} \min_{v, \gamma, Z^1, Z^2} \; & f(v, \gamma, Z^1, Z^2) \\ \text{s. t. } & \langle v^{\ell-1}, v\rangle = 1, \ (33), \ (34), \ (36). \end{aligned} \tag{41}$$

Denote its solution by $\hat{\xi}^\ell = (\hat{v}^\ell, \hat{\gamma}^\ell, \hat{Z}^{1,\ell}, \hat{Z}^{2,\ell})$.
Step 2 If $f(\hat{\xi}^\ell) = f(\xi^{\ell-1})$, then stop; otherwise set

$$v^\ell = \frac{1}{\|\hat{v}^\ell\|}\hat{v}^\ell, \quad \gamma^\ell = \frac{1}{\|\hat{v}^\ell\|}\hat{\gamma}^\ell, \quad Z^{i,\ell} = \frac{1}{\|\hat{v}^\ell\|}\hat{Z}^{i,\ell}, \ i = 1, 2.$$

Increase ℓ by one and go to Step 1.

Theorem 7 *When the method stops, the point $(Z^{1,\ell}, Z^{2,\ell}, v^\ell, \gamma^\ell)$ satisfies the optimality conditions for problem (38). Otherwise, the method generates a sequence of points $\big\{(Z^{1,\ell}, Z^{2,\ell}, v^\ell, \gamma^\ell)\big\}_{\ell=1}^{\infty}$ such that every accumulation point satisfies the optimality conditions for problem (38).*

Proof First, we formulate optimality conditions for problem (38). Assign Lagrange multipliers $\mu^1 \in \mathbb{R}^{m_1}_+$, $\mu^2 \in \mathbb{R}^{m_2}_+$ and $\mu^3 \in \mathbb{R}$ to the constraints (33) (re-formulated to $\leq$), (34), and (35), respectively. The Lagrange function has the form

$$\begin{aligned} \Lambda(v, \gamma, Z^1, Z^2) &= \lambda\varrho_1(Z^1) + (1-\lambda)\varrho_2(Z^2) + \sum_{i=1}^{m_1} \mu_i^1(-\langle v, x_j^1\rangle + \gamma - z_i^1) \\ &\quad + \sum_{j=1}^{m_1} \mu_j^2(\langle v, x_j^2\rangle - \gamma - z_j^2) + \mu^3(\langle v, v\rangle - 1) \\ &= \lambda\varrho_1(Z^1) + (1-\lambda)\varrho_2(Z^2) - \sum_{i=1}^{m_1} \mu_i^1 z_i^1 - \sum_{j=1}^{m_2} \mu_j^2 z_j^2 \\ &\quad + \sum_{i=1}^{m_1} \mu_i^1(-\langle v, x_j^1\rangle + \gamma) + \sum_{j=1}^{m_1} \mu_j^2(\langle v, x_j^2\rangle - \gamma) + \mu^3(\langle v, v\rangle - 1). \end{aligned}$$

If the point $(\bar{v}, \bar{\gamma}, \bar{Z}^1, \bar{Z}^2)$ is optimal, then we have

$$0 = \nabla_{v,\gamma} \Lambda(\bar{v}, \bar{\gamma}, \bar{Z}^1, \bar{Z}^2); \tag{42}$$

$$0 \in \partial_{Z^1, Z^2} \Lambda(\bar{v}, \bar{\gamma}, \bar{Z}^1, \bar{Z}^2) + \mathcal{N}_{\mathbb{R}_+^{m_1+m_2}}(\bar{Z}^1, \bar{Z}^2), \tag{43}$$

$$\mu_i^1(-\langle v, x_j^1\rangle + \gamma - z_i^1) = 0, \quad i = 1, \ldots, m_1, \tag{44}$$

$$\mu_j^2(\langle v, x_j^2\rangle - \gamma - z_j^2) = 0, \quad j = 1, \ldots, m_2. \tag{45}$$

Condition (42) has the form

$$-\sum_{i=1}^{m_1} \mu_i^1 x_i^1 + \sum_{i=1}^{m_2} \mu_i^2 x_i^2 + 2\mu^3 \bar{v} = 0 \tag{46}$$

$$\sum_{i=1}^{m_1} \mu_i^1 = \sum_{i=1}^{m_2} \mu_i^2. \tag{47}$$

Condition (43) is equivalent to the existence of subgradients $\zeta^1 \in \partial \varrho_1(\bar{Z}^1)$ and $\zeta^2 \in \partial \varrho_2(\bar{Z}^2)$ such that

$$\begin{aligned} &\lambda\zeta^1 \geq \mu^1 \geq 0, \quad \langle \lambda\zeta^1 - \mu^1, \bar{z}^1\rangle = 0, \\ &(1-\lambda)\zeta^2 \geq \mu^2 \geq 0, \quad \langle (1-\lambda)\zeta^2 - \mu^2, \bar{z}^2\rangle = 0. \end{aligned} \tag{48}$$

It is easy to see that, for $\bar{v} = v^{\ell-1}$, then conditions (46)–(47)–(48) coincide with the optimality conditions for problem (41) at iteration ℓ with optimal Lagrange multipliers μ^1, μ^2 and $2\mu^3 \in \mathbb{R}$.

If the method stops at iteration ℓ, we have $f(\hat{\xi}^\ell) = f(\xi^{\ell-1})$. The point $\xi^{\ell-1}$ is feasible for problem (41) at iteration ℓ. Therefore, $\xi^{\ell-1}$ is optimal for (41) and satisfies the optimality conditions for (41). Since $\|v^{\ell-1}\| = 1$, the point $\xi^{\ell-1}$ is feasible for problem (38). Therefore, the optimality conditions for (38) are also satisfied at ξ^ℓ.

Now consider the case when the method generates an infinite sequence of points $\{\xi^\ell\}$. For any solution $\hat{\xi}^\ell$ of problem (41), we have

$$1 = \langle v^{\ell-1}, \hat{v}^\ell\rangle \leq \|v^{\ell-1}\|.\|\hat{v}^\ell\| = \|\hat{v}^\ell\|.$$

Thus, if the method does not stop at iteration ℓ, the following inequality holds:

$$f(\xi^\ell) = \frac{1}{\|\hat{v}^\ell\|} f(\hat{\xi}^\ell) < f(\hat{\xi}^\ell) < f(\xi^{\ell-1}).$$

Consequently, the sequence $\{f(\xi^\ell)\}$ is monotonically decreasing.

Since $\|v^\ell\| = 1$, then γ^ℓ, as well as $Z^{1,\ell}$, $Z^{2,\ell}$ are bounded (cf. (37), (29)). Thus, the sequence $\{\xi^\ell\}$ has a convergent subsequence $\mathscr{L} \subset \{1, 2, \ldots\}$. Let $\xi^* = (v^*, \gamma^*, Z_*^1, Z_*^2)$ be an accumulation point of $\{\xi^\ell\}$. Due to the continuity of f and the monotonicity of $\{f(\xi^\ell)\}$, we have

$$\lim_{\ell \to \infty} f(\xi^\ell) = f(\xi^*).$$

Furthermore, the point ξ^* is feasible for (38) because all points ξ^ℓ are feasible and the constraint functions are continuous.

Each point $\hat{\xi}^\ell$ satisfies the optimality conditions for problem (41). Let $\zeta^{1,\ell} \in \partial \varrho_1(\hat{Z}^{1,\ell})$ and $\zeta^{2,\ell} \in \partial \varrho_2(\hat{Z}^{1,\ell})$ be the optimal subgradients from the corresponding condition (48). Due to the positive homogeneity of the risk measures, it holds

$$\zeta^{1,\ell} \in \partial \varrho_1(Z^{1,\ell}) \quad \text{and} \quad \zeta^{2,\ell} \in \partial \varrho_2(Z^{2,\ell}).$$

Using the fact that the subdifferential mapping of a convex function is upper semi-continuous with compact values, we obtain that the sequence $\{(\zeta^{1,\ell}, \zeta^{2,\ell})\}$, $\ell \in \mathscr{L}$, has a convergent subsequence $\mathscr{L}_1 \subset \mathscr{L}$ by virtue of Berge theorem, i.e. $\lim_{\ell \in \mathscr{L}_1}(\zeta^{1,\ell}, \zeta^{2,\ell}) = (\zeta^1_*, \zeta^2_*)$. This implies that the optimal Lagrange multipliers $\mu^{1,\ell}, \mu^{2,\ell}$ are bounded due to the inequalities (48) in the optimality conditions. Therefore, the optimal Lagrange multipliers $\{\mu^{1,\ell}, \mu^{2,\ell}\}$ are convergent to μ^1_*, μ^2_* for a subsequence $\mathscr{L}_2 \subset \mathscr{L}_1$. Finally, $\mu^{3,\ell}$ is convergent to some number μ^3_* for $\ell \in \mathscr{L}_2$, which is obtained passing to the limit in Eq. (46). We conclude that the point ξ^* satisfies the optimality conditions for (38) with Lagrange multipliers $\mu^1_*, \mu^2_*, \mu^3_*$ and subgradients (ζ^1_*, ζ^2_*). □

7 Related work

The design of robust estimators, robust classifiers in particular, has attracted attention of statisticians as well as of data scientists. Misclassification may lead to different cost distribution for the different types of errors. An example illustrating this point is the damage caused by a hurricane. The cost of the hurricane damage depends on features, which are used for classification. The cost is highly non-linear with respect to those features (see Davis and Uryasev 2016). Therefore, if we fail to predict correctly that a hurricane will take place in a certain region, the cost is quite different than the cost induced by an incorrect hurricane alarm. Furthermore, different predictions of the hurricane's location lead to different cost. A risk-averse model prediction would take this fact into account (see, e.g. Chambers and Quiggin 2000). Another example is classification of credit-worthiness of bank customers. If a customer, who might be a company requesting a substantial loan, is classified incorrectly as credit worthy, the bank may experience substantial loss while not providing a loan to a credit-worthy customer, results in a lost opportunity; both losses are quite different Oguz et al. (2008)

Different attitude to errors in model fitting was proposed a long time ago in statistics and this point of view is accepted and used in various approaches, most notably, in robust statistics. We refer to Huber (2011), El Ghaoui et al. (2003), Gotoh and Uryasev (2017), Hastie et al. (2009) and the references therein for methods of robust classification design for binary classification. Support vector machines are one of the most popular classification tools. They appear as part of sequential classification methods for multiple classes Sculley et al. (2011). Recent developments include the use of SVMs as part of deep learning architectures (Kim et al. 2015; Qi et al. 2016). Other recent papers (Zareapoor et al. 2018), leverage the power of Deep Belief Networks as input to SVMs in ensemble algorithms. For information on kernel methods in classification and support vector learning, we refer to Schölkopf et al. (1999), Maji et al. (2008), Muandet et al. (2012). Additionally, there is a substantial research effort into incremental learning for SVM and Support Vector Regression (Liang and Li 2009; Gu et al. 2015a, b).

Many papers deal with robust binary classification. One possibility is provided by the tool of robust statistics; for example, employing the Huber risk function (Huber 2011). We refer to Zhang (2004) for detailed discussion on robustness and choice of loss functions. In Lanckriet et al. (2003) and El Ghaoui et al. (2003), the tools of robust optimization are employed. The idea there is that the future instance will come from a distribution, which is close to the observed empirical distribution in some sense. Therefore, a set of acceptable distributions is constructed, called an uncertainty set, and the worst misclassification error

is minimized over all distributions in that set. In Lanckriet et al. (2003) and El Ghaoui et al. (2003), the uncertainty sets are defined by allowing all distributions on the sample space, which have the same mean and the same covariance as the estimated ones. In Ma et al. (2011) the authors look at the median rather than the expected value and the sum of the two median errors is minimized. In Katsumata and Takeda (2015), the authors allow for uncertainty sets of different diameter for each class.

Our proposed approach suggests to minimize the classification error in a risk averse manner. In Rockafellar et al. (2008), the use of coherent measures of risk for generalized regression and model fit was proposed. This point of view was also utilized in SVM in Gotoh and Uryasev (2017). While those works recognize the need of expressing different attitude to errors in fitting statistical models, the authors propose using one overall measure of risk as an objective in the regression problem, respectively in the SVM problem. The classification design based on a single measure of risk does not allow for differentiation between the classes, while in our view different attitude should be allowed to classification errors for the different classes.

The topic of risk sharing is a subject of intensive investigations in the community of economics, quantitative finance and risk management. This is due to the fact that the sum of the risk of each component in a system does not equal the risk of the entire system. The main focus in the extant literature on risk-sharing is on the choice of decomposition of a random variable X into k terms $X = X^1 + \cdots + X^k$, so that when each component is measured by a specific risk measure, the associated total risk is in some sense optimal. The variable X represents the total random loss of the firm and the question addressed is about splitting the loss among the constituents. Assigning coherent measures of risk ϱ_i to each term X^i, the adopted point of view is that the outcome $\big(\varrho_1(X^1), \ldots, \varrho_k(X^k)\big)$ should be Pareto-optimal among the feasible allocations.

The main results in risk-sharing theory accomplish the decomposition of X into terms by looking at the infimal convolution of the measures of risk. It is observed (see, e.g., Landsberger and Meilijson 1994; Ludkovski and Rüschendorf 2008) that the random variables X^i, $i = 1, \ldots, k$, which solve this problem, satisfy a co-monotonicity property as follows

$$\big(X^i(\omega) - X^i(\omega')\big)\big(X^j(\omega) - X^j(\omega')\big) \geq 0, \quad \text{for all } \omega, \omega' \in \Omega, \quad i, j = 1, \ldots, k.$$

While we adopt similar point of view on optimality of risk allocation, it is clear that the problem setting and subsequently the results associated with risk sharing in financial institutions are inapplicable to the classification problem. We cannot expect co-monotonicity properties of the class errors because not all decomposition of the total random error can be obtained via some classifier. The presence of constraints in the optimization problem, the functional dependence of the misclassification error on the classifier's parameters, and the complex nature of design problem require dedicated analysis.

8 Numerical experiments

In the previous sections, we have shown the solid theoretical foundation supporting our approach. In this section, we display the performance of the proposed framework, as well as its flexibility. To this end, we use several publicly available data sets and compare the performance of our approach to some existing formulations, in terms of performance metrics frequently used in classification. Further, we showcase the flexibility of the framework by exploring the Pareto-efficient frontier of various classifiers derived from our framework.

Table 1 Data summary

Data set	Features	Observations		Class
		Class0	Class1 (%)	Balance
WDBC	30	357	211 (37.1)	0.591
pima-indians-diabetes	7	500	267 (34.8)	0.534
seismic-bumps	18	2414	170 (6.6)	0.070

In our numerical experiments, we have used the Conditional Value-at-Risk and the mean semi-deviation of order one.

8.1 Data

We compare our approach to other known approaches on several datasets. More specifically, we use three data sets obtained from the UCI Machine Learning Repository (Lichman 2013). These data sets exhibit different degrees of class imbalance, that is the proportion of records in one class versus that of the other class. A summary of basic characteristics of the data sets is shown in Table 1.

In the sections to follow, we make reference to the default and target class for each dataset. This nomenclature is consistent with the group represented by the class itself. In other words, the default class (Class0) in the WDBC dataset contains the observations where the diagnosis was "benign", while the target class (Class1) represents observations with a "malignant" diagnosis. Similarly, for the other two datasets, the default class represents the healthy or normal state while the target class represents the class of interest in the context of the data. For the "pima-indians-diabetes" dataset the target class is the set diabetics amongs the sample, while for the "seismic-bumps" dataset the target class is the set of shifts where high energy seismic bumps occurred.

8.2 Model formulations

We consider several scenarios for choices of measures of risk. In the first scenario, we treat the default class (Class0) in a risk neutral manner, while applying the mean-semi-deviation measure to the classification error of the target class. We call this loss function "asym_risk" (see Table 2). In the same table, we provide the risk measure combinations for other loss functions which we have used in our numerical experiments. The loss functions called "risk_cvar" and "two_cvar" use a convex combination of the expected error and the Conditional Value-at-Risk of the classification error. These convex combinations use an additional model parameter $\beta \in (0, 1)$. The formulation (38) for these loss functions uses the variational form of the Conditional Value-at-Risk at level $\alpha \in (0, 1)$. Table 2 displays the chosen combinations of risk measure pairs for the binary classification scenario in order to give an easy overview.

We note that calculation of the first order semi-deviation and the conditional value-at-risk can be formulated as linear optimization problems. Therefore, their application does not increase the complexity of RSSVM in comparison to the soft-margin SVM. However, if we use higher order semi-deviations or higher order inverse risk measures, the problem becomes more difficult.

Table 2 Risk measure combinations used as loss functions in the experiments

Loss function	Class0—$\varrho_1(Z^1)$	Class1—$\varrho_2(Z^2)$
exp_val	$\mathbb{E}[Z^1]$	$\mathbb{E}[Z^2]$
joint_cvar	$\beta\mathbb{E}[Z^1+Z^2]+(1-\beta)$	$\text{CVaR}_\alpha(Z^1+Z^2)$
asym_risk	$\mathbb{E}[Z^1]$	$\mathbb{E}[Z^2]+c\sigma^+[Z^2]$
one_cvar	$\mathbb{E}[Z^1]+c\sigma^+[Z^1]$	$\text{CVaR}_\alpha(Z^2)$
risk_cvar	$\mathbb{E}[Z^1]+c\sigma^+[Z^1]$	$\beta\mathbb{E}[Z^2]+(1-\beta)\text{CVaR}_\alpha(Z^2)$
two_risk	$\mathbb{E}[Z^1]+c\sigma^+[Z^1]$	$\mathbb{E}[Z^2]+c\sigma^+[Z^2]$
two_cvar	$\beta\mathbb{E}[Z^1]+(1-\beta)\text{CVaR}_{\alpha_1}(Z^1)$	$\beta\mathbb{E}[Z^2]+(1-\beta_2)\text{CVaR}_{\alpha_2}(Z^2)$

Note that conditional value-at-risk at level α puts weight on the largest quantiles of the error distribution (the upper α-portion of them). This means the classifier is focused mainly on eliminating the worst classification errors. Depending on our risk-aversion, we could include small or larger portion of quantiles by controlling the level α. Additionally, taking convex combinations of the expected value and the conditional value-at-risk (using $\beta > 0$), we include into consideration all quantiles with some additional weight on the largest quantile. The higher the value of β the less weight it is put on the largest misclassification errors and $\beta = 1$ corresponds to the risk-neutral SVM. The mean-semideviation measure adds weight to all classification errors which are above average size. As already mentioned, higher value of the constant c entails larger penalty for deviations above the mean. Furthermore, using higher order semi-deviations results in a non-linear (form of power function) penalty for those deviations.

We compare our results against three different benchmarks: two risk-neutral formulations and one risk-averse formulation with a single risk measure. The first risk-neutral formulation is the soft-margin SVM as formulated in (4). The second risk-neutral formulation uses the Huber loss function and leads to the following problem formulation

$$\min_{v,\gamma,Z^1,Z^2}\left\{\frac{1}{m_1}\sum_{i=1}^{m_1}\min\left(z_i^1,(z_i^1)^2\right)+\frac{1}{m_2}\sum_{j=1}^{m_2}\min\left(z_j^2,(z_j^2)^2\right)\text{: (33), (34), (36)}\right\}. \tag{49}$$

The third benchmark uses a single risk measure (50) on the total error as proposed in Gotoh and Uryasev (2017). It has the following formulation.

$$\begin{aligned}\min_{v,\gamma,t,Z^1,Z^2,Y^1,Y^2}\ & \beta\left(\frac{1}{m_1}\sum_{j=1}^{m_1}z_j^1+\frac{1}{m_2}\sum_{j=1}^{m_2}z_j^2\right)\\ & +(1-\beta)\left(t+\frac{1}{\alpha(m_1+m_2)}\left(\sum_{j=1}^{m_1}y_j^1+\sum_{j=1}^{m_2}y_j^2\right)\right)+\delta\|v\|^2\\ \text{s. t. }\ & y_j^i\geq z_j^i-t,\quad j=1,\ldots,m_i,\ i=1,2,\\ & \text{(33), (34), (36)},\ Y^1\geq 0,\quad Y^2\geq 0.\end{aligned} \tag{50}$$

Interestingly, both risk-neutral formulations produce nearly identical results on all data sets. Subsequently we only report one of them under the name "exp_val". In the presented figures and tables below, we refer to the loss function consisting of a single Conditional Value-at-Risk measure, as "joint_cvar".

The problem formulations used in our experiments are the following.

Expected value vs. Conditional Value-at-Risk—"asym_risk"

$$\begin{aligned} \min_{v,\gamma,t,Z^1,Z^2,Y} \quad & \frac{\lambda}{m_1}\sum_{j=1}^{m_1} z_j^1 + \frac{1-\lambda}{m_2}\sum_{j=1}^{m_2}(y_j + z_j^2) \\ \text{s. t.} \quad & y_j \geq z_j^2 - t, \quad j = 1,\ldots,m_2, \\ & (33), (34), (35), (36), \ Y \geq 0. \end{aligned} \tag{51}$$

Mean-semi-deviation vs. Conditional Value-at-Risk—"one_cvar"

$$\begin{aligned} \min_{v,\gamma,t,Z^1,Z^2,Y^1,Y^2} \quad & \frac{\lambda}{m_1}\sum_{j=1}^{m_1}(y_j^1 + z_j^1) + (1-\lambda)\left(t + \frac{1}{\alpha m_2}\sum_{j=1}^{m_2} y_j^2\right) \\ \text{s. t.} \quad & y_j^1 \geq z_j^1 - \frac{1}{m_1}\sum_{j=1}^{m_1} z_j^1, \quad j = 1,\ldots,m_1, \\ & y_j^2 \geq z_j^2 - t, \quad j = 1,\ldots,m_2, \\ & (33), (34), (35), (36), \ Y^1 \geq 0, \quad Y^2 \geq 0. \end{aligned} \tag{52}$$

Mean-semi-deviation vs. combination of the expectation and CVaR—"risk_cvar"

$$\begin{aligned} \min_{v,\gamma,t,Z^1,Z^2,Y^1,Y^2} \quad & \frac{\lambda}{m_1}\sum_{j=1}^{m_1}(y_j^1 + z_j^1) + \frac{\beta(1-\lambda)}{m_1}\sum_{j=1}^{m_2} z_j^2 \\ & + (1-\beta)(1-\lambda)\left(t + \frac{1}{\alpha m_2}\sum_{j=1}^{m_2} y_j^2\right) \\ \text{s. t.} \quad & y_j^1 \geq z_j^1 - \frac{1}{m_1}\sum_{j=1}^{m_1} z_j^1, \quad j = 1,\ldots,m_1, \\ & y_j^2 \geq z_j^2 - t, \quad j = 1,\ldots,m_2, \\ & (33), (34), (35), (36), \ Y^1 \geq 0, \quad Y^2 \geq 0. \end{aligned} \tag{53}$$

Mean-semi-deviation for both classes—"two_risk"

$$\begin{aligned} \min_{v,\gamma,Z^1,Z^2,Y^1,Y^2} \quad & \frac{\lambda}{m_1}\sum_{j=1}^{m_1}(y_j^1 + z_j^1) + \frac{1-\lambda}{m_2}\sum_{j=1}^{m_2}(y_j^2 + z_j^2) \\ \text{s. t.} \quad & y_j^i \geq z_j^i - \frac{1}{m_i}\sum_{j=1}^{m_i} z_j^i, , \quad j = 1,\ldots,m_i, \ i = 1,2, \\ & (33), (34), (35), (36), \ Y^1 \geq 0, \ Y^2 \geq 0. \end{aligned} \tag{54}$$

Table 3 Main results table for the WDBC dataset: displaying the model parameters and the performance metrics for each model formulation

exp_val		joint_cvar	asym_risk	one_cvar	risk_cvar	two_risk	two_cvar
F_1-score optimized classifiers							
lambda			0.70	0.57	0.56	0.60	0.64
alpha_1							0.62
alpha_2		0.55		0.88	0.75		0.62
C0 errors	21	17	16	13	11	15	12
C1 errors	15	11	11	10	9	9	9
FPR	0.05882	0.04762	0.04482	0.03641	0.03081	0.04202	0.03361
Recall	0.92925	0.94811	0.94811	0.95283	0.95755	0.95755	0.95755
Precision	0.90367	0.92202	0.92627	0.93953	0.94860	0.93119	0.94419
F_1-score	0.91628	0.93488	0.93706	0.94614	**0.95305**	0.94419	0.95082
AUC	0.97904	0.98426	0.98569	**0.98764**	0.98535	0.98442	0.98451
AUC optimized classifiers							
lambda			0.43	0.57	0.69	0.37	0.42
alpha_1							0.61
alpha_2		0.65		0.88	0.66		0.61
C0 errors	21	21	18	13	14	23	16
C1 errors	15	13	11	10	13	12	13
FPR	0.05882	0.05882	0.05042	0.03641	0.03922	0.06443	0.04482
Recall	0.92925	0.93868	0.94811	0.95283	0.93868	0.94340	0.93868
Precision	0.90367	0.90455	0.91781	0.93953	0.93427	0.89686	0.92558
F_1-score	0.91628	0.92130	0.93271	**0.94614**	0.93647	0.91954	0.93208
AUC	0.97904	0.98471	0.98697	0.98764	0.98776	0.98629	**0.98922**

The boldface numbers indicate the best performing model formulation (column) with respect to the specified performance metric (row). In other words, there is is only one bold face number per row for each of the performance metrics, F_1-score and AUC

Conditional-Value at Risk for both classes—"two_cvar"

$$\begin{aligned} \min_{v,\gamma,t_1,t_2,Z^1,Z^2,Y^1,Y^2} \quad & \lambda\beta_1 \sum_{j=1}^{m_1} z_j^1 + \lambda(1-\beta_1)\left(t_1 + \frac{1}{\alpha m_1}\sum_{j=1}^{m_1} y_j^1\right) \\ & + (1-\lambda)\beta_2 \sum_{j=1}^{m_1} z_j^2 + (1-\lambda)(1-\beta_2)\left(t_2 + \frac{1}{\alpha m_2}\sum_{j=1}^{m_2} y_j^2\right) \qquad (55) \\ \text{s. t.} \quad & y_j^i \geq z_j^i - t_i, \quad j = 1,\ldots,m_i,\ i = 1,2, \\ & (33), (34), (35), (36),\ Y^1 \geq 0,\ Y^2 \geq 0. \end{aligned}$$

8.3 Performance

We perform k-fold cross-validation and all reported results are out of sample. Furthermore, our method determines only normalized optimal classifiers and all misclassification numbers and reported risk are computed with respect to such classifiers. In Tables 3, 5, and 7, we

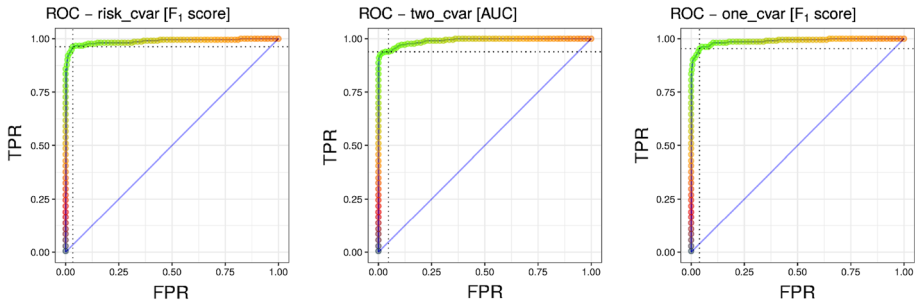

Fig. 2 ROC plots for the best performing model formulations on the WDBC data: "risk_cvar" with the best F_1-score, "two_cvar" with the best AUC value, and "one_cvar" for the alternate metric

report the F_1-score and AUC, along with recall (True Positive Rate), precision, as well as false positive rate (FPR) for all loss functions. Additionally, we report the number of misclassified observations, as well as the chosen parameters, where applicable. In light of the fact that the F_1-score and AUC are competing metrics, for each dataset we present two set of results, one optimized for each metric. We use this to highlight the additional flexibility offered by the proposed method as discussed in the next section.

We recall that precision is the ratio of true positives over the total number of positively classified points. F_1-score, or sometimes referred to as F–measure, is the harmonic mean between precision and recall.

$$\text{F}_1\text{-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

AUC stands for Area Under the [Receive Operating Characteristic] Curve, and is defined as the integral of the curve created by the True Positive Rate and False Positive Rate as functions of the threshold.

$$\text{AUC} = \int_{-\infty}^{\infty} \text{TPR}(T)\text{FPR}(T)dT$$

In Table 3, we show the best value for each metric for each set in bold face. We observe that for this particular dataset, the best performing model formulation with respect to the F_1-score is the "risk_cvar" model; outperforming the risk neutral formulations by more than 0.04. On the other hand, if we consider the AUC to be the target metric, we notice the "two_cvar" formulation has the highest value. Further, we note that the "one_cvar" model has the same parameters for both target metrics. We find this to be unusual in our experiments. While this formulation does not have the best value for the target metric, it too significantly outperforms the risk neutral formulations. Further, this formulation does have the best value for the competing metric in both cases. The respective ROC curves for each of the classifiers are displayed in Fig. 2. The color on each curve represents the value of the F_1-score. High values are represented by the bright green color, and low values are represented by the dark red color. The two dotted lines indicate the threshold at which the classifier is set to operate.

We can certainly see the classifier performs very well on this data. Table 4 contains the calculations of risk, with respect to each model formulation. More specifically, for each obtained classifier we calculate the value of the risk functionals on the out of the sample data points during cross-validation. We consider the raw expectation, mean semi-deviation, as well as the conditional value-at-risk for the α quantiles 0.75, 0.85, and 0.95.

Table 4 Risk evaluation for the WDBC data set: displaying the expectation of error, mean semi-deviation, and average value at risk for the α quantiles 0.75, 0.85, and 0.95

	Expectation	MSD	$\text{CVaR}_{0.75}$	$\text{CVaR}_{0.85}$	$\text{CVaR}_{0.95}$
WDBC					
exp_val					
C0 risk	0.000189	0.000368	0.000252	0.000223	0.000199
C1 risk	0.000343	0.000663	0.000457	0.000403	0.000361
Total	0.000532	0.001030	0.000709	0.000626	0.000560
joint_cvar					
C0 risk	0.000158	0.000309	0.000211	0.000186	0.000167
C1 risk	0.000241	0.000470	0.000322	0.000284	0.000254
Total	0.000400	0.000779	0.000533	0.000470	0.000421
asym_risk					
C0 risk	0.000121	0.000237	0.000161	0.000142	0.000127
C1 risk	0.000194	0.000378	0.000259	0.000228	0.000204
Total	0.000315	0.000615	0.000420	0.000371	0.000332
one_cvar					
C0 risk	0.000085	0.000166	0.000113	0.000100	0.000089
C1 risk	0.000172	0.000335	0.000229	0.000202	0.000181
Total	0.000256	0.000501	0.000342	0.000302	0.000270
risk_cvar					
C0 risk	0.000080	0.000157	0.000106	0.000094	0.000084
C1 risk	0.000185	0.000363	0.000247	0.000218	0.000195
Total	0.000265	0.000520	0.000353	0.000312	0.000279
two_risk					
C0 risk	0.000125	0.000246	0.000167	0.000148	0.000132
C1 risk	0.000182	0.000356	0.000242	0.000214	0.000191
Total	0.000307	0.000601	0.000410	0.000361	0.000323
two_cvar					
C0 risk	0.000085	0.000167	0.000113	0.000100	0.000089
C1 risk	0.000235	0.000460	0.000314	0.000277	0.000248
Total	0.000320	0.000628	0.000427	0.000377	0.000337

Indeed, we can observe that our models reduce the risk for each class with respect to each risk calculation, compared to the benchmarks. More specifically, we notice that the "one_cvar" model, which does not attain the best performance in terms of F_1-score, but does, in fact, attain the lowest total risk value. Its value is approximately one half that of the risk neutral formulation, and that of the other benchmark. The "risk_cvar" model does perform nearly identically, albeit having at slightly larger values across the board. Further, we note that the "two_cvar" model, which performs best with respect to the AUC metric is the worst performing, benchmarks excluded. Looking closely at the corresponding ROC curve in Fig. 2 one can argue that the performance with respect to the AUC metric, comes at the expense of robustness and generalization.

Looking at the results on the "pima-indians-diabetes" data set in Table 5 we observe that the best performing model with respect to F_1-score is the again "risk_cvar" model with 0.68581

Table 5 Main results table for the "pima-indians-diabetes" dataset: displaying the model parameters and the performance metrics for each model formulation

exp_val		joint_cvar	asym_risk	one_cvar	risk_cvar	two_risk	two_cvar
F_1-score optimized classifiers							
lambda			0.48	0.51	0.49	0.48	0.44
alpha_1							0.58
alpha_2		0.90		0.68	0.56		0.58
C0 errors	107	92	158	121	125	129	157
C1 errors	80	93	46	65	62	63	48
FPR	0.21400	0.18400	0.31600	0.24200	0.25000	0.25800	0.31400
Recall	0.70149	0.65299	0.82836	0.75746	0.76866	0.76493	0.82090
Precision	0.63729	0.65543	0.58421	0.62654	0.62236	0.61377	0.58355
F_1-score	0.66785	0.65421	0.68519	0.68581	**0.68781**	0.68106	0.68217
AUC	0.83039	0.83243	0.82900	**0.83078**	0.83033	0.82967	0.82830
AUC optimized classifiers							
lambda			0.51	0.54	0.54	0.50	0.60
alpha_1							0.69
alpha_2		0.59		0.86	0.76		0.69
C0 errors	107	87	140	80	79	113	74
C1 errors	80	98	59	99	99	78	106
FPR	0.21400	0.17400	0.28000	0.16000	0.15800	0.22600	0.14800
Recall	0.70149	0.63433	0.77985	0.63060	0.63060	0.70896	0.60448
Precision	0.63729	0.66148	0.59885	0.67871	0.68145	0.62706	0.68644
F_1-score	0.66785	0.64762	**0.67747**	0.65377	0.65504	0.66550	0.64286
AUC	0.83039	0.83279	0.83081	**0.83348**	0.83332	0.83049	0.83267

The boldface numbers indicate the best performing model formulation (column) with respect to the specified performance metric (row). In other words, there is is only one bold face number per row for each of the performance metrics, F_1-score and AUC

compared to the 0.66785 of the risk neutral formulations. Similarly, the "one_cvar" model is again second in this context, at the same time having the largest AUC value for the group. Surprisingly, the benchmark formulation "joint_cvar" has the lowest score here. Switching the attention to the AUC section of the table, we notice that "one_cvar" is the best performing model in that regard as well; with the "risk_cvar" being second best. However, the gain in AUC value with the changed parameters is minimal with a considerable reduction in the alternate target metric; "one_cvar" shifting from 0.68581 F_1-score to 0.65377 in exchange for 0.0027 gain in AUC, and "risk_cvar" shifting from 0.68781 F_1 to 0.65504 for a gain of 0.003.

Figure 4 shows how the empirical distribution of error realizations from applying the classifier to out-of-sample records on the left, and the overlayed ROC curves for the various classifiers on the right. Negative values indicate correctly classified observations, while positive values indicate misclassification. We compare the select loss functions to each other and the benchmarks. Looking closely at the ROC curves in Fig. 3, we can see that the AUC prioritized "one_cvar" actually does not classify at its maximum potential in terms of F_1-score, indicated by the fact that the threshold is not at the lightest green segment of the curve. This requires additional investigation and exploration. The shape of the ROC curves for the various classifiers is relatively similar, see right panel of Fig. 4. However, looking

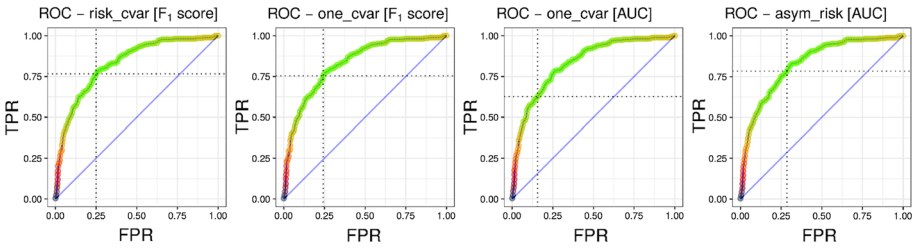

Fig. 3 ROC plots for the best performing model formulations on the "pima-indians-diabetes" data: "risk_cvar" with the best F_1-score, "one_cvar" featuring both parameter sets, and finally the "asym_risk" formulation featuring the best AUC value

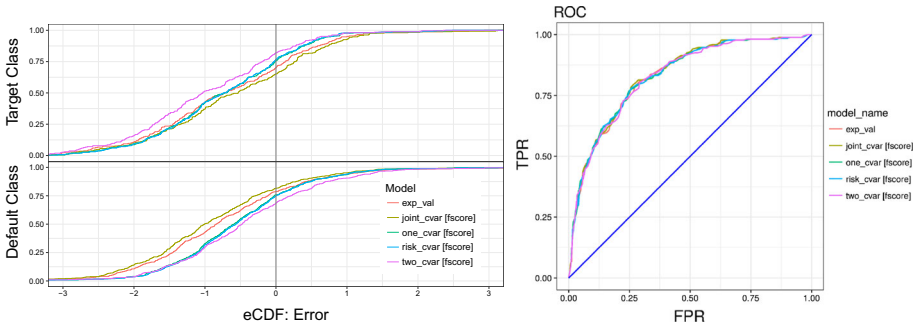

Fig. 4 Empirical distribution of error realizations comparing risk-averse loss function formulations to benchmarks [F_1-score] on the "pima-indians-diabetes" dataset (left) and the corresponding ROC curves (right)

at the error distribution plot on the left in the figure, we notice that the two benchmarks, "exp_val" and "join_cvar", misclassify less of the default class and more of the target class. It is important to show the similarity of the ROC curves in order to highlight how the choice of risk measure may impact the convergence to a specific threshold. More specifically, we see the "two_cvar" formulation underperforming, in relation to the target metric (F_1-score) and the best performing formulation "risk_cvar", by misclassifying too much of the default class (Table 6).

Table 7 contains the risk functional evaluation for the "pima-indians-data". It is interesting that the "two_risk" model has the lowest total risk with respect to every risk functional, despite the fact that is not the best performing model in terms of F_1-score or AUC. This leads us to believe that there may be room for additional exploration with regard to performance metrics and evaluation.

We continue with the performance evaluation on the third and final dataset, whose main performance metrics are shown in Table 7. One can immediately observe, that no model performs particularly well on this dataset. We have chosen this data set for being particularly imbalanced and containing categorical variables.

Again, we see the "risk_cvar" formulation as having the best F_1-score, followed very closely by the "joint_cvar" formulation. In terms of AUC, it is the "two_cvar" formulation that leads group, but again at a significant cost of the F_1-score. Looking at Fig. 5, we can see room for improvements to the this by changing the threshold on the AUC prioritized "two_cvar" model. We observe that in terms of stability to that respect, the "asy_risk" formulation along

Table 6 Risk evaluation for the "pima-indians-diabetes" data set: displaying the expectation of error, mean semi-deviation, and average value at risk for the α quantiles 0.75, 0.85, and 0.95

	Expectation	MSD	$\text{CVaR}_{0.75}$	$\text{CVaR}_{0.85}$	$\text{CVaR}_{0.95}$
pima-indians-diabetes					
exp_val					
C0 risk	0.164317	0.296266	0.219089	0.193314	0.172965
C1 risk	0.183513	0.318461	0.244684	0.215898	0.193172
Total	0.347830	0.614727	0.463773	0.409212	0.366137
joint_cvar					
C0 risk	0.132718	0.242794	0.176957	0.156138	0.139703
C1 risk	0.226791	0.383421	0.302387	0.266812	0.238727
Total	0.359508	0.626215	0.479344	0.422951	0.378430
asym_risk					
C0 risk	0.251054	0.431147	0.334738	0.295357	0.264267
C1 risk	0.092539	0.169554	0.123385	0.108869	0.097409
Total	0.343593	0.600701	0.458124	0.404227	0.361676
one_cvar					
C0 risk	0.167050	0.296830	0.222733	0.196529	0.175842
C1 risk	0.128815	0.229708	0.171754	0.151547	0.135595
Total	0.295865	0.526538	0.394487	0.348077	0.311437
risk_cvar					
C0 risk	0.168882	0.299515	0.225176	0.198685	0.177771
C1 risk	0.123088	0.220300	0.164118	0.144810	0.129567
Total	0.291970	0.519815	0.389294	0.343495	0.307337
two_risk					
C0 risk	0.152290	0.269093	0.203053	0.179165	0.160305
C1 risk	0.110126	0.195772	0.146835	0.129560	0.115922
Total	0.262416	0.464865	0.349888	0.308725	0.276227
two_cvar					
C0 risk	0.240685	0.415233	0.320913	0.283158	0.253352
C1 risk	0.103057	0.188842	0.137409	0.121244	0.108481
Total	0.343742	0.604075	0.458322	0.404402	0.361833

with "joint_cvar" benchmark have less variation. Turning the attention to the risk functional evaluation in Table 8, we observe that the "exp_val" benchmark model has the lowest total on the "seismic-bumps". However, being that this dataset is very imbalanced, we can see how significantly different the risk functional evaluation is between the two classes for each model formulation.

Notice, in Fig. 6, how the "exp_val" benchmark stands alone compared to the well grouped risk aware models, which includes the benchmark formulation "joint_cvar". Similarly, as on the previous dataset, the ROC curves are very much grouped.

In summary, the F_1-score prioritized model consistently provides small but significant improvement over the baseline models.

Table 7 Main results table for the "seismic-bumps" dataset: displaying the model parameters and the performance metrics for each model formulation

exp_val		joint_cvar	asym_risk	one_cvar	risk_cvar	two_risk	two_cvar
F_1-score optimized classifiers							
lambda			0.61	0.60	0.59	0.53	0.70
alpha_1							0.92
alpha_2		0.60		0.86	0.84		0.92
C0 errors	471	203	269	248	230	270	201
C1 errors	64	93	83	85	87	83	94
FPR	0.19511	0.08409	0.11143	0.10273	0.09528	0.11185	0.08326
Recall	0.62353	0.45294	0.51176	0.50000	0.48824	0.51176	0.44706
Precision	0.18371	0.27500	0.24438	0.25526	0.26518	0.24370	0.27437
F_1-score	0.28380	0.34222	0.33080	0.33797	**0.34369**	0.33017	0.34004
AUC	0.76157	0.75482	**0.76187**	0.75595	0.75496	0.75133	0.75629
AUC optimized classifiers							
lambda			0.60	0.47	0.47	0.49	0.47
alpha_1							0.56
alpha_2		0.93		0.75	0.58		0.56
C0 errors	471	261	292	812	817	571	633
C1 errors	64	84	82	50	48	62	54
FPR	0.19511	0.10812	0.12096	0.33637	0.33844	0.23654	0.26222
Recall	0.62353	0.50588	0.51765	0.70588	0.71765	0.63529	0.68235
Precision	0.18371	0.24784	0.23158	0.12876	0.12993	0.15906	0.15487
F_1-score	0.28380	**0.33269**	0.32000	0.21779	0.22002	0.25442	0.25245
AUC	0.76157	0.76068	0.76360	0.76489	0.76611	0.76344	**0.76637**

The boldface numbers indicate the best performing model formulation (column) with respect to the specified performance metric (row). In other words, there is is only one bold face number per row for each of the performance metrics, F_1-score and AUC

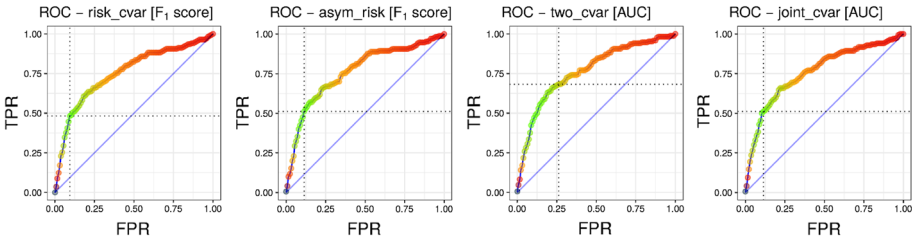

Fig. 5 ROC plots for the best performing model formulations on the "seismic-bumps" data: "risk_cvar" with the best F_1-score, "one_cvar", "joint_cvar", "two_cvar" formulation featuring the best AUC value

8.4 Flexibility

Our approach provides additional flexibility which is generally not available for classification methods like soft-margin SVM. We allow the user to implement a predetermined attitude toward risk of misclassification, and to explore the Pareto-efficient frontier of classifiers.

Table 8 Risk Evaluation for the "seismic-bumps" data set: displaying the expectation of error, mean semi-deviation, and conditional value at risk for the α quantiles 0.75, 0.85, and 0.95

	Expectation	MSD	$CVaR_{0.75}$	$CVaR_{0.85}$	$CVaR_{0.95}$
seismic-bumps					
exp_val					
C0 risk	0.039589	0.072043	0.052786	0.046576	0.041673
C1 risk	0.064462	0.106979	0.085950	0.075838	0.067855
Total	0.104052	0.179022	0.138735	0.122414	0.109528
joint_cvar					
C0 risk	0.015641	0.030007	0.020854	0.018401	0.016464
C1 risk	0.131682	0.199685	0.175576	0.154920	0.138613
Total	0.147323	0.229693	0.196430	0.173321	0.155077
asym_risk					
C0 risk	0.018930	0.035855	0.025239	0.022270	0.019926
C1 risk	0.099935	0.156758	0.133246	0.117570	0.105194
Total	0.118864	0.192613	0.158485	0.139840	0.125120
one_cvar					
C0 risk	0.019387	0.036922	0.025850	0.022809	0.020408
C1 risk	0.116238	0.179858	0.154983	0.136750	0.122355
Total	0.135625	0.216780	0.180833	0.159559	0.142763
risk_cvar					
C0 risk	0.015942	0.030445	0.021256	0.018755	0.016781
C1 risk	0.107669	0.164839	0.143559	0.126669	0.113336
Total	0.123611	0.195284	0.164814	0.145424	0.130116
two_risk					
C0 risk	0.015797	0.029943	0.021062	0.018584	0.016628
C1 risk	0.088633	0.139315	0.118177	0.104274	0.093298
Total	0.104430	0.169258	0.139239	0.122858	0.109926
two_cvar					
C0 risk	0.013332	0.025589	0.017776	0.015685	0.014034
C1 risk	0.110821	0.167536	0.147762	0.130378	0.116654
Total	0.124153	0.193126	0.165538	0.146063	0.130688

We traverse the Pareto frontier by varying λ from 0.4 to 0.7 and observe that the solution is rather sensitive to the scalarization used in the loss function. In Fig. 7, we show the resulting error densities from such a traversal. We can observe how varying the weight between the two risk measures allows us to obtain a family of risk-averse Pareto-optimal classifiers. The efficient frontier can be used to choose a risk-averse classifier according *additional* criterion as the F_1-score, AUC, or other similar performance metrics by choosing specific parameter λ, as discussed in the previous section.

The Pareto frontier looks substantially different when different combinations of risk measures are used. Further research would reveal the effect of higher order risk measures and their ability to create a classifier with highly discriminant powers.

We have chosen the probability level for the Conditional Value-at-Risk in a similar way. We observe that the loss function "one_cvar" consistently provides the best performance.

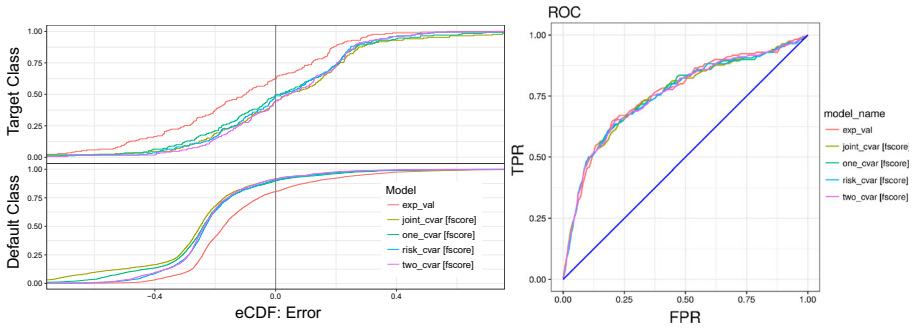

Fig. 6 Empirical distribution of error realizations comparing risk-averse loss function formulations to benchmarks [F_1-score] on the "seismic-bumps" dataset (left) and the corresponding ROC curves (right)

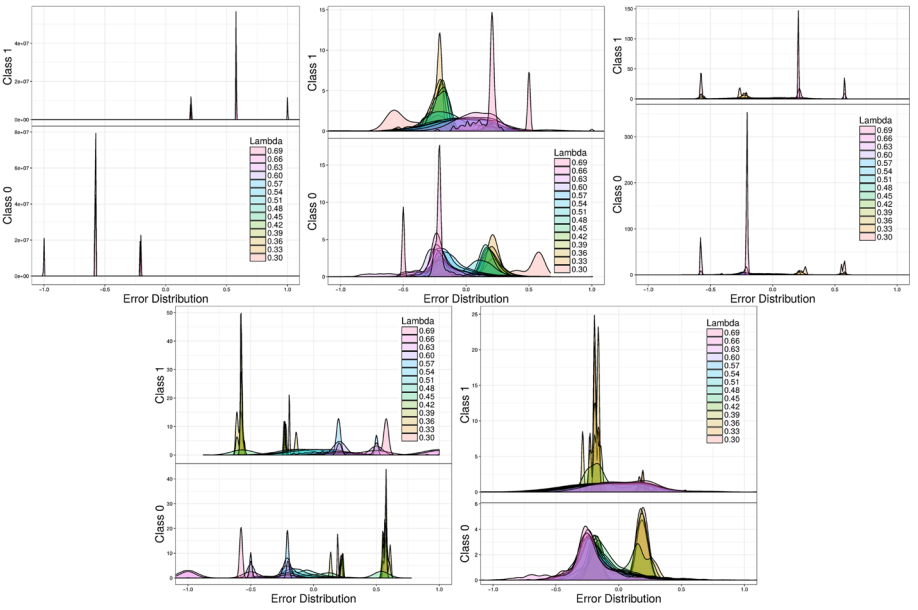

Fig. 7 The distribution of error displayed as smoothed histogram for each of five proposed formulations for the risk-averse SVM problem e.g. "asym_risk", "one_cvar", "risk_cvar", "two_risk", and "two_cvar" all using the same set of λ values, with other parameters fixed, on the "seismic-bumps" dataset

A close second, is the loss function "risk_cvar," which has a similar structure. Interestingly, using the same risk measure on both classes does not perform as well.

Acknowledgements The authors thank the anonymous referees whose vigorous criticism helped improve the paper.

References

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*(3), 203–228.

Beer, G. (1993). *Topologies on closed and closed convex sets* (Vol. 268). Berlin: Springer.
Chambers, R. G., & Quiggin, J. (2000). *Uncertainty, production, choice, and agency: The state-contingent approach*. Cambridge: Cambridge University Press.
Cox, D. R., & Lewis, P. A. W. (1966). *The statistical analysis of series of events*. London: Methuen.
Davis, J. R., & Uryasev, S. (2016). Analysis of tropical storm damage using buffered probability of exceedance. *Natural Hazards*, *83*(1), 465–483.
Dentcheva, D., & Martinez, G. (2012). Two-stage stochastic optimization problems with stochastic ordering constraints on the recourse. *European Journal of Operational Research*, *219*(1), 1–8.
Dentcheva, D., & Penev, S. (2010). Shape-restricted inference for Lorenz curves using duality theory. *Statistics & Probability Letters*, *80*(5), 403–412.
Dentcheva, D., Penev, S., & Ruszczyński, A. (2010). Kusuoka representation of higher order dual risk measures. *Annals of Operations Research*, *181*(1), 325–335.
Dentcheva, D., Penev, S., & Ruszczyński, A. (2017). Statistical estimation of composite risk functionals and risk optimization problems. *Annals of the Institute of Statistical Mathematics*, *69*(4), 737–760.
El Ghaoui, L., Lanckriet, G. R. G., Natsoulis, G., et al. (2003). *Robust classification with interval data*. Berkeley: Computer Science Division, University of California.
Fano, U. (1947). Ionization yield of radiations. II. The fluctuations of the number of ions. *Physical Review*, *72*(1), 26.
Föllmer, H., & Schied, A. (2011). *Stochastic finance: An introduction in discrete time*. Berlin: Walter de Gruyter.
Gotoh, J., & Uryasev, S. (2017). Support vector machines based on convex risk functions and general norms. *Annals of Operations Research*, *249*(1–2), 301–328.
Gu, B., Sheng, V. S., Tay, K. Y., Romano, W., & Li, S. (2015a). Incremental support vector learning for ordinal regression. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(7), 1403–1416.
Gu, B., Sheng, V. S., Wang, Z., Ho, D., Osman, S., & Li, S. (2015b). Incremental learning for ν-support vector regression. *Neural Networks*, *67*, 140–150.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction. Springer series in statistics* (2nd ed.). New York: Springer.
Huber, P. J. (2011). *Robust statistics*. Berlin: Springer.
Jouini, E., Schachermayer, W., & Touzi, N. (2008). Optimal risk sharing for law invariant monetary utility functions. *Mathematical Finance*, *18*(2), 269–292.
Katsumata, S., & Takeda, A. (2015). Robust cost sensitive support vector machine. In *Artificial intelligence and statistics* (pp. 434–443).
Kijima, M., & Ohnishi, M. (1993). Mean-risk analysis of risk aversion and wealth effects on optimal portfolios with multiple investment opportunities. *Annals of Operations Research*, *45*(1), 147–163.
Kim, S., Yu, Z., Kil, R. M., & Lee, M. (2015). Deep learning of support vector machines with class probability output networks. *Neural Networks*, *64*, 19–28.
Krokhmal, P. A. (2007). Higher moment coherent risk measures. *Quantitative Finance*, *7*(4), 373–387.
Kusuoka, S. (2001). On law invariant coherent risk measures. In *Advances in mathematical economics* (pp. 83–95). Tokyo: Springer.
Lanckriet, G. R. G., El Ghaoui, L., Bhattacharyya, C., & Jordan, M. I. (2003). A robust minimax approach to classification. *The Journal of Machine Learning Research*, *3*, 555–582.
Landsberger, M., & Meilijson, I. (1994). Co-monotone allocations, Bickel–Lehmann dispersion and the Arrow–Pratt measure of risk aversion. *Annals of Operations Research*, *52*(2), 97–106.
Liang, Z., & Li, Y. F. (2009). Incremental support vector machine learning in the primal and applications. *Neurocomputing*, *72*(10–12), 2249–2258.
Lichman, M. (2013). UCI machine learning repository.
Ludkovski, M., & Rüschendorf, L. (2008). On comonotonicity of pareto optimal risk sharing. *Statistics & Probability Letters*, *78*(10), 1181–1188.
Ma, Y., Li, L., Huang, X., & Wang, S. (2011). Robust support vector machine using least median loss penalty. In *the 16th IFAC world congress* (pp. 11208–11213).
Maji, S., Berg, A. C., & Malik, J. (2008). Classification using intersection kernel support vector machines is efficient. In *IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008* (pp. 1–8). IEEE.
Miettinen, K. (1999). *Nonlinear multiobjective optimization* (Vol. 12). Berlin: Springer.
Muandet, K., Fukumizu, K., Dinuzzo, F., & Schölkopf, B. (2012). Learning from distributions via support measure machines. In *Advances in neural information processing systems* (pp. 10–18).
Ogryczak, W., & Ruszczyński, A. (1999). From stochastic dominance to mean-risk models: Semideviations as risk measures. *European Journal of Operational Research*, *116*(1), 33–50.

Ogryczak, W. L., & Ruszczyński, A. (2002). Dual stochastic dominance and related mean risk models. *SIAM Journal on Optimization*, *13*(1), 60–78.

Oguz, H. T., & Gurgen, F. S. (2008). Credit risk analysis using hidden Markov model. In *23rd International symposium on computer and information sciences, 2008. ISCIS'08*. IEEE.

Qi, Z., Wang, B., Tian, Y., & Zhang, P. (2016). When ensemble learning meets deep learning: A new deep support vector machine for classification. *Knowledge-Based Systems*, *107*, 54–60.

Rockafellar, R. T., & Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, *26*(7), 1443–1471.

Rockafellar, R. T., Uryasev, S., & Zabarankin, M. (2006). Generalized deviations in risk analysis. *Finance and Stochastics*, *10*(1), 51–74.

Rockafellar, R. T., Uryasev, S., & Zabarankin, M. (2008). Risk tuning with generalized linear regression. *Mathematics of Operations Research*, *33*(3), 712–729.

Ruszczynski, A., & Shapiro, A. (2006). Optimization of convex risk functions. *Mathematics of Operations Research*, *31*(3), 433–452.

Schölkopf, B., Burges, C. J. C., & Smola, A. J. (Eds.). (1999). *Advances in kernel methods: Support vector learning*. MIT press.

Sculley, D., Otey, M. E., Pohl, M., Spitznagel, B., Hainsworth, J., & Zhou, Y. (2011). Detecting adversarial advertisements in the wild. In *Proceedings of the 17th ACM SIGKDD international conference on data mining and knowledge discovery*.

Shaked, M., & Shanthikumar, J. G. (2007). *Stochastic orders*. Springer.

Shapiro, A., Dentcheva, D., & Ruszczyński, A. (2009). *Lectures on stochastic programming: Modeling and theory*. Philadelphia: SIAM.

Shapiro, A., Dentcheva, D., & Ruszczyński, A. (2014). *Lectures on stochastic programming: Modeling and theory* (Vol. 16). Philadelphia: SIAM.

Zareapoor, M., Shamsolmoali, P., Jain, D. K., Wang, H., & Yang, J. (2018). Kernelized support vector machine with deep learning: An efficient approach for extreme multiclass dataset. *Pattern Recognition Letters*, *115*, 4–13.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, *32*(1), 56–85.