TOWARDS RISK MODELS IN MACHINE LEARNING

By

CONSTANTINE ALEXANDER VITT

A dissertation submitted to the

Graduate School—Newark

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Information Technology

written under the direction of

Hui Xiong

and approved by

_____

_____

_____

_____

Newark, New Jersey

May, 2018

ABSTRACT OF THE DISSERTATION

Towards Risk Models in Machine Learning

By CONSTANTINE ALEXANDER VITT

Dissertation Director:

Hui Xiong

This thesis explores new models for some machine learning problems based on recent developments in the theory and methods of risk analysis and risk-averse optimization. Two types of risk models are used: coherent measures of risk and a dynamic model based on stochastic differential equations. These models are applied to two areas of machine learning: classification and identification of the impact of patent activity on the stock-price dynamics of companies in the technology sector.

We propose a new approach to classification, which aims at determining a risk-averse classifier. It allows different attitude to misclassification risk for the different classes. This is accomplished by the application of non-linear risk functions specific to each class. The structure of the new classification problems is analyzed and optimality conditions are obtained. We show that the risk-averse classification problem is equivalent to an optimization problem with unequal, implicitly defined but unknown

weights for each data point. The new methodology is implemented in a binary classification scenario in several versions. One type of risk-averse SVM is based on a soft-margin classifier using various coherent measures of risk as objectives. Another type of risk-averse SVM problems determine a classifier with a normalized vector of the separation plane using again several sets of risk measures in the respective objectives. We propose a numerical method for solving the classification problem with normalization constraint. Numerical test are performed on several data sets with different levels of separation difficulty. The results are compared to classification with benchmark loss functions, which are well established in the literature.

In the second part of the thesis, we consider patent activity in the technology sector and their impact on the stock-price dynamics. We show the promises of exploiting patent data for the analysis and prospecting of high-tech companies in the stock market. A new approach to analyze the relationships between patent activities and statistical characteristics of the stock price is developed, which may be of interest to discovery of statistical relations among sequential data beyond this context. We demonstrate the relationships between the monthly drift and volatility of the market adjusted stock returns and the number of patent applications as well as the diversity of the corresponding patent categories. We use a widely accepted model of the market-adjusted stock returns and estimate its parameters. Adopting the moving window technique, we fit models by introducing various lagged terms of patent activity characteristics. For each company, we consider the coefficients of each significant term over the entire time horizon and perform further statistical hypothesis testing on the overall significance of the corresponding indicator. The analysis has been performed on real-world stock trading data as well as patent data. The results confirm the impact of innovations on stock movement and show that the market-adjusted stock

returns do exhibit more volatility if the company has been extending their patents to new areas. On the other hand, the statistical relation between the drift of stock returns and the patent activity of a company appears to be of more complex nature involving other latent factors.

Acknowledgments

I would like to express my gratitude to my advisor Prof. Xiong for the support and encouragement of my Ph.D study, as well as his patience and motivation.

I would like to thank Prof. Dentcheva who extended support, through her immense knowledge and guidance, without which it would not be possible to conduct this research. I am grateful to Prof. Ruszczyński for the valuable and inspiring discussions.

I would like to thank Prof. Lin, Prof. Kuang, and Prof. Papadimitriou, who graciously served on my dissertation committee, providing valuable and insightful comments.

I would like to thank the entire Management Science & Information Systems Department and the Research Group for the seminars and research discussions.

The financial support from the Rutgers Business School is gracefully acknowledged.

Dedication

I would like to dedicate this thesis to my parents, whose love, unselfish support, encouragement, and guidance made this all possible; and to my grandparents who helped shape me into the person I am today, being always there for me, even from across the world.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

Due to the rapid development of technology in recent years, large sets of observational data have become available in almost all areas of science, engineering, and business. This creates the opportunity of analyzing data and extracting information, which opens new perspectives, scientific discoveries, as well as business advantages for the successful explorers. This opportunity has motivated the recent interest in data mining.

Most data sets are gathered from dynamical systems whose evolution is of interest. In some cases, we seek to discover typical patterns of evolution, which are inherent for the systems. We may be interested in predicting the future behavior, or we look for anomalies in the current behavior. The theoretical and numerical challenges associated with the analysis of large data sets have attracted the attention of many highly qualified scientists who have developed a number of new algorithms to classify, cluster, segment, index, model, and detect anomalies in them.

In every specific problem arising in machine learning, we deal with a given dataset, which we may want to represent by mapping each instance to a space of features. The features may be continuous, categorical or binary and are supposed to express the essence of the information contained in each record. Feature selection is an important process in which, we identify the most relevant features of each instance in the dataset to our task; we remove redundant features or transform the instances to obtain new (transformed) features. This process reduces the dimensionality of the data and

facilitates the operation of data mining algorithms. We talk about supervised learning when the instances are given with known labels assigning them to some categories or classes. In contrast to this, unsupervised learning deals with unlabeled instances.

Typical techniques of machine learning are the following.

- **Classification**: It is the most notable technique of supervised learning. Given a task and a dataset, the goal is to identify a function, which will output an proper assignment to any future instance to one of two or more predefined classes [38, 50, 54].

- **Clustering**: Data clustering is one of the most popular data labeling techniques. Its goal is to determine task-appropriate groupings of the instances in a database under some similarity (or dissimilarity) measure $D(\cdot, \cdot)$ [45, 59, 78, 50]. Clustering can be applied to an exploratory task or as a preprocessing step for further machine learning work.

- **Indexing (Query by Content)**: Given a query X, and some similarity (or dissimilarity) measure $D(\cdot, \cdot)$, identify the most similar instance in a database [25, 70, 80, 49].

- **Motif Discovery** is the detection of previously unknown, frequently occurring patterns (see, e.g.,[21, 101]);

- **Statistical analysis**: several typical tasks include **segmentation** of sequential data into sections of stationary processes with similar characteristics, create an approximation of X which retains its essential features but is of much smaller size (also known as **summarization**); most notably **Prediction** of future values and **Anomaly detection**. Anomaly may be considered "unexpected / unusual / novelty" occurrences, which may lead to a big risk or to high profit. This task is also associated with outliers detections [5] or rare events discovery [18].

In all of these tasks, we see the essential role of a proper statistical model and a good way to identify patterns of interest. That is why, in this thesis, we present a contribution to those two areas. In the first part of the thesis, we focus of classification problems and develop a risk-averse approach to such problems. In the second part of the thesis, we focus on a dynamical system and propose a new methodology for gaining insight into its dynamics.

Classification is one of the most frequently used data-mining method. It is a fundamental tool used in anomaly detection, serving to detect fraud, equipment failure, insider trading, health anomalies, security bridges, e-mail spam and phishing, etc. In our numerical work, we have used data to identify health anomaly and patterns associated with seismic activities pertaining to mining and safety.

The problem of detecting anomalous, unusual and or novel patterns has been a point of focus for many researchers. The problem of anomaly detection consists in defining what constitutes an anomalous pattern. In order to do this, we must first determine what is considered a normal state or pattern. Next, we must ascertain what magnitude of deviation from this normal state is to be considered abnormal. Online detection of abnormal patterns/subsequence requires efficient treatment of dynamic data steam, availability of data for training and validation purposes, striking a good balance between rapid detection and low false alarm rate. Additionally, we need to monitor and identify changes in the boarders of normal and abnormal regions. The available data may be noisy or the system may be subject to changes. These challenges are known in statistical analysis and many scientist have suggested analytical and numerical tools to address them to some extent. We refer to the surveys [66, 17, 46] for systematic presentation of the anomaly detection techniques. Classification based anomaly detection techniques are developed on the basis of binary separation, as well as multi-class separation. A binary classifier is a function on the space of features with a binary output. Given an observation, the task is to classify it as anomalous, in which case, the classifier outputs one after reading it, or normal; the output of

the classifier is zero. This approach requires a training set, in which the normal sequences are known/labeled. In many applications, multiple normal classes can be identified. In that context, a binary classifier is trained to distinguish between each normal class and the rest of the classes. Each subsequent instance is considered anomalous if it is not classified as normal by any of the classifiers. Other application (e.g., [91]) associates multiple classes of anomalous (malicious) rare events which have different nature. A new instance is considered normal, if it is not classified as malicious (anomalous) of any kind. The theoretical analysis in this thesis will cover those type of classifications. The methodology proposed here contributes to the stability and robustness of the classification by involving the modern tools of coherent measures of risk. Additionally, we provide a method to construct confidence intervals of the risk associated with the new type of classifiers. This is a novel information supplied in addition to the confidence score currently associated with the prediction made by the classifier.

Time series data accounts for an increasingly large fraction of the world's supply of data. A random sample of 4,000 graphics from 15 of the world's newspapers published from 1974 to 1989 found that more than 75% of all graphics were time series [100]. Time-series data mining addresses a wide range of real-life tasks in various fields of research. Some examples include economic forecasting [32], intrusion detection [15, 13], gene expression analysis [84], medical surveillance [14], hydrology [77], and virtually all areas of human activity. In the medical domain alone, large streams of data such as gene expression data [11], electrocardiograms, electroencephalograms, growth development charts are routinely recorded and analyzed. A lot of interest and attention attracts data related to finance, as well as data related to entertainment, meteorology, and many other industries. In finance, a new research area, called high-frequency data analysis, has emerged [35], which seeks to extract information from almost continuous data streams. On-line identification of specific rare events for high-frequency data in the context of quantitative finance is discussed in [10].

We shall outline a specific application of time series data-mining techniques in finance, where we analyze what impact the patent activity of a high-tech company has on the dynamics of its stock.

In this thesis, we have used publicly available data with an appropriate preprocessing. The raw recorded data coming from dynamical systems are usually high-dimensional streams which are often recorded with errors; they may contain data gaps, erroneous records, duplicate records, or disordered sequences. All these problems specifically occurred in relation to the problem presented in chapter 3. Some type of erroneous records are relatively easy to address and tools for their removal are available. The procedures to mitigate these errors are referred to as "data cleaning", but, to the best or our knowledge, no formal definition of this process exists. Additionally, no clear delineation between data cleaning and data quality exists. Usually the existing literature suggests to combine several methods in order to increase the power of detecting errors in the data. In many cases, gaps are present in the data series. Several methods are available for dealing with this problem. The simplest method is to use statistical estimates such as a mean or a median value instead of a missing value; such estimates can be calculated on the basis of the neighboring values (cf. [42, 30, 61]). It is accepted that good practice for data management require proper documentation of all procedures associated with it. Data cleaning is as an essential aspect of quality assurance and it is important when one seeks to compare algorithms or validate the studies using the data. We discuss the specific data preprocessing methods germane to our study in due course.

Chapter 2
Risk Sharing

2.1 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 proliferates throughout the business world, engineering, and the sciences. Our goal is to propose a new approach to classification problems and to develop a methodology for reliable risk-averse classifiers design which has the flexibly to allow customers choice of risk measurement for the misclassification errors in various classes. 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. The theoretical results on risk sharing are inapplicable here. The classification problem raises new issues, poses new challenges, and requires a dedicated analysis.

We consider labeled data consisting of k subsets $S_1, \ldots, S_k$ of n-dimensional vectors. The labels of the data points in S_i will be denoted by y_i and $y_i = i$. The cardinality of S_i is $|S_i| = m_i$, $i = 1, \ldots, k$. The data points represent observations in the space of "features" (i.e., the number of features is n). Analytically, classification problem consists in identifying a mapping φ, whose image can be partitioned into k subsets corresponding to each class of data, so that $\varphi(\cdot)$ can be used as an indicator function of each class. We adopt the following definition.

Definition 1. *A classifier is a function* $\varphi : \mathbb{R}^n \to \mathbb{R}^d$ *and a collection of nonempty*

sets $K_i \subset \mathbb{R}^d$, $i = 1, \ldots, k$, *such that*

$$\varphi(x) \in K_i \ \ \textit{for all } x \in S_i,\ i = 1, \ldots, k,$$
$$K_i \cap K_j = \emptyset \ \textit{for all } i, j = 1, \ldots, k,\ i \neq j.$$

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 values for the parameter π, that is, we deal with a vector valued classifier $\varphi(x; \pi) = (\varphi_1(x; \pi), \ldots, \varphi_d(x; \pi))^\top$ and regions $K_i \in \mathbb{R}^d$, $i = 1, \ldots, k$. Some examples of this point of view are the following.

Example 1 (**The support vector machine**)**.**
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 $\varphi(x; \pi) = v^\top x - \gamma$ 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 $K_1 = [0, +\infty)$, $K_2 = (-\infty, 0)$.

Example 2 (**Polyhedral classifier for multiple classes**)**.**
Let us consider the case of separating many classes, e.g., $k \geq 3$ by the creating a linear classifier on the principle "one vs. all". Then effectively, our goal is to determine functions $\varphi_j(x; a^j, b_j) := \langle a^j, x \rangle - b_j$, where x is a data point from the feature space, $a^j \in \mathbb{R}^n$, $j = 1, \ldots k-1$, are the normals of the separating planes and b_j determine the location of the j-th plane. Plane j is meant to separate the data points from class j from the rest of the data points. This means that

$$\varphi_j(x; a^j, b_j) = \begin{cases} \geq 0 & \text{for } x \in S_j \\ < 0 & \text{for } x \notin S_j. \end{cases} \tag{2.1}$$

We define a $k - 1 \times n$ matrix A whose rows are the vectors a^j, and a vector $b \in \mathbb{R}^{k-1}$ whose components are b_j. 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 $\pi = (A, b) \in \mathbb{R}^{(k-1)(n+1)}$. Requirement (2.1) means that the regions K_j are the orthants

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

This setting may be used for classification in the anomaly detection scenario. Two approaches are possible. 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.

Example 3 (**Kernel-based classifier**)**.**
Kernel methods for classification assume that the data is mapped to a high dimensional space of features where the classes of data are more likely to be separable. That space must be a pre-Hilbert space, i.e., a space $\mathcal{Z}$, where inner product is defined and a reproducing kernel $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ exists. More precisely, the non-linear mapping ψ is such that $\psi : \mathbb{R}^n \to \mathcal{Z}$

$$K(x, x') = \langle \psi(x), \psi(x') \rangle_{\mathcal{Z}},$$

where $\langle \cdot, \cdot \rangle_{\mathcal{Z}}$ denotes the inner product in $\mathcal{Z}$. The function ψ is defined implicitly by the choice of the kernel. In this case, we transfer our analysis to the new feature space $\mathcal{Z}$: we determine the regions K_i, $i = 1, \ldots k$ as being part of $\mathcal{Z}$ and consider classifiers φ defined on $\mathcal{Z}$. We intend to provide more precise formulation in due course.

2.2 Loss Functions

A key element, which distinguishes various classification approaches, is the choice of a loss function, which combines several goals. On the one hand, it serves as a

model fitting loss function in the statistical sense while minimizing misclassification. On the other hand, it also controls model flexibility and numerical stability when the classification problem is solved. Typically, the loss function is chosen as one of the known risk functionals in statistical model fitting. The quality of every model is determined by analysis of the residuals, e.g. the error. Let us introduce the following notation. 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 $i = 1,\dots,k$. In statistical terms, we try to predict the membership $y \in \{1,\dots,k\}$ of a data point to one of the classes. Therefore, we try to determine $\varphi(x;A,b)$ in such a way that the probability of the following events is maximized

$$\mathbb{P}\Big\{\bigcap_{i=1}^{k}\big(\varphi(x;\pi) \in K_i | x \in S_i\big)\Big\}.$$

Alternatively, we are interested in minimizing the probability of the event

$$\mathbb{P}\Big\{\bigcup_{i=1}^{k}\big(\varphi(x;\pi) \notin K_i | x \in S_i\big)\Big\}.$$

Using the indicator function of an event, we can estimate the aforementioned probability as follows:

$$\mathbb{P}\big(\varphi(x;\pi) \notin K_i | x \in S_i\big) = \frac{1}{m_i}\sum_{x_j \in S_i} \bar{\mathbb{I}}_{K_i}(x_j),$$

$$\text{where } \bar{\mathbb{I}}_{K_i}(x) = \begin{cases} 0 & \text{if } \varphi(x;\pi) \in K_i \\ 1 & \text{if } \varphi(x;\pi) \notin K_i. \end{cases}$$

The classification error can be defined as the distance of a particular record to the classification set, to which it should belong. Here the distance from a point r to a set K is defined by using a suitable norm in $\mathbb{R}^n$:

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

Note that here we assumes implicitly that the set K is convex and closed. Indeed, the sets K_i, $i = 1\dots,k$ are closed convex set for most classification problems, as evidenced by the examples in the previous section.

In statistical terms, 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 random variables defined as follows

$$Z^i(\pi) = \text{dist}(\varphi(X^i; \pi), K_i), \;\; i = 1, \ldots k, \tag{2.2}$$

represent the misclassification of records in class i when parameter π is used. These are univariate random variables defined on the same probability space and are represented by the sampled observations

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

The expected misclassification error for each class can be estimated as follows:

$$\hat{Z}^i(\pi) = \sum_{x_j \in S_i} \frac{1}{m_i} \text{dist}(\varphi(x_j; \pi), K_i)$$

The following figure illustrates how the classification error for a certain binary classifier is measured. In our examples, we consider the distance of the points $\varphi(x; \pi)$ to

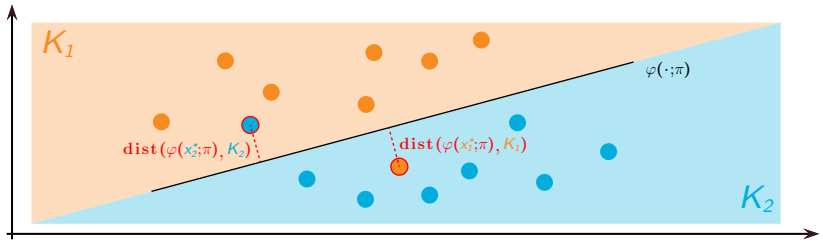

Figure 2.1: Classification error calculation

the sets K_i, $i = 1, \ldots, k$

Example 4 (The support vector machine continued).

In the support vector machine the classification error is computed by

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

We classify every new observation x in S_i, if 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 dist $\big(\varphi(x; v, \gamma), K_i\big) = 0$, $i, = 1, 2$, always holds for exactly one class.

Example 5 (**Polyhedral classifier for multiple classes continued**).
Observe that in this case, the regions K_i, $i = 1, \dots k$ do not cover the entire image space of the classifier. Therefore, it is possible to to observe a future instance x such that dist $\big(\varphi(x; A, b), K_i\big) > 0$ for all $i = 1, \dots k$. In that case, we could classify according to the smallest distance

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

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.

Example 6 (**Kernel based classifiers continued**).
If we have chosen a kernel with associated mapping ψ such that

$$K(x, x') = \langle \psi(x), \psi(x') \rangle_{\mathcal{Z}},$$

then the relevant distances become dist $\Big(\varphi\big(\psi(x); \pi\big), K_i\Big)$, $i = 1, \dots k$. We shall continue the analysis of this example for binary classification in section 2.9.

2.3 Robust Classification Design and Robust Statistics

The design of robust estimators, robust classifiers in particular, has attracted attention of statisticians as well as of data scientists. Additionally, the distributions of the populations providing the currently available records may not be well represented by the current sample (e.g., it might have heavy tails, not be unimodal, etc.) Furthermore, misclassification may lead to different cost with different probability depending

on the error. An example for such a case is the damage caused by a hurricane. If we fail to predict correctly that a hurricane will take place in certain region, the cost of the damage depends on the features used for classification and is highly non-linear with respect to those features (see [24]).

We refer to [43, 31, **?**, 40] and the references therein for methods of robust classification design. Most cases address binary classification.

Support vector machines are widely used and most popular classification tools. They appear also as part of sequential classification methods for multiple classes. Various approaches in the literature address the design of a robust classifier specifically for the support vector machine.

We start with the formulation of an optimization problem based on the loss function expressing the minimization of the (estimated) expected total classification error. The design of a binary classifier can be accomplished by solving the following optimization problem:

$$\begin{aligned} \min_{v,\gamma,Z^1,Z^2} \ & \frac{1}{m_1}\sum_{j=1}^{m_1} z_j^1 + \frac{1}{m_2}\sum_{j=1}^{m_2} z_j^2 \\ \text{s. t.} \ & \langle v, x_j^1\rangle - \gamma + z_j^1 \geq 0, \quad j = 1,\ldots,m_1, \\ & \langle v, x_j^2\rangle - \gamma - z_j^2 \leq 0, \quad j = 1,\ldots,m_2, \\ & \|v\| = 1, \\ & Z^1 \geq 0, Z^2 \geq 0. \end{aligned} \tag{2.3}$$

In this formulation, Z^1 and Z^2 are random variables expressing the magnitude of the classification error for class 1 and class 2, respectively. Those variables have realizations z_i^1 and z_i^2. The parameters of the classifier are $\pi = (v, \gamma)$. Note that proper calculation of the magnitude of classification error requires the use of the Euclidean norm of v. In that case, $\langle v, x\rangle = \gamma$ is the equation of a plane and the value $\varphi(x;\pi) = \langle v, x\rangle - \gamma$ is indicative of the position of the point x relative to that plane: the sign of $\varphi(x;\pi)$ indicates on which side of the plane the point is located and the absolute value of $\varphi(x;\pi)$ indicates how far is the point from the plane.

The presence of the constraint $\|v\| = 1$ makes problem (2.3) non-convex and, therefore, difficult to solve. The problem is frequently replaced by the so-called soft-margin SVM, which is formulated as follows:

$$\begin{aligned} \min_{v,\gamma,Z^1,Z^2} \ & \frac{1}{m_1}\sum_{j=1}^{m_1} z_j^1 + \frac{1}{m_2}\sum_{j=1}^{m_2} z_j^2 + \delta\|v\|^2 \\ \text{s. t.} \ & \langle v, x_j^1\rangle - \gamma + z_j^1 \geq 1, \quad j = 1,\dots,m_1, \\ & \langle v, x_j^2\rangle - \gamma - z_j^2 \leq -1, \quad j = 1,\dots,m_2, \\ & Z^1 \geq 0, Z^2 \geq 0. \end{aligned} \tag{2.4}$$

Here $\delta > 0$ is a small number. The objective function of problem (2.4) uses additional regularization term consisting of the squared norm of the normal vector to the separating hyperplane. In problem (2.4), the normal vector v can be of any positive length. Note that $v = 0$ would result in large error and it is, therefore, not a candidate for the optimal solution. Additionally, the regularization term $\delta\|v\|^2$ prevents vector v to become too large in the optimal solution. Observe that multiplying the solution of problem (2.4), v and γ, by a positive constant does not change the separating plane. In problem (2.4), 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. Therefore, it only makes sense to compare the quality of normalized classifiers, where the length of v is one.

Most notable approach to robust binary classification is provided by the theory and methods of robust statistics. In this approach, the model is fit using the Huber risk function, which is defined for $z \in S_i$, $i = 1, 2$ as follows:

$$L_H(z; v, \gamma) = \begin{cases} \left[\max\left(0, 1 + (-1)^i(\gamma - \langle v, z\rangle)\right)\right]^2 & \text{if} \\ \qquad\qquad (-1)^i(\gamma - \langle v, z\rangle) \geq -1 & \\ (-1)^i(\langle v, z\rangle - \gamma) \qquad \text{otherwise.} & \end{cases} \tag{2.5}$$

Another approach is presented in [57, 31], where the tools of robust optimization are employed. The idea there is that the future instance will come a distribution, which is close to the observed empirical distribution in some sense. Therefore, some set of distrributions is constructed, called uncertainty set, and the minimization is carried out over all distributions in that set. In [57, 31], the uncertainty sets are defined by allowing all distributions on the smape space, which have the same mean and the same covariance as the estimated empirical mean and covariance. In [67] the authors look at the median hinge loss determined for each class and minimize the sum of the two median losses.

Our proposed approach suggests to minimize the classification error in a risk averse manner. For this purpose, we propose new family of loss functions, which use coherent measures of risk.

2.4 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 [34] or the index of dispersion [22]) are of similar spirit.

An axiomatic theory of measures of risk is presented in [75, 3, 36, 51, 85] In a more general setting risk measures are analyzed in [90]. 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. We use $\overline{\mathbb{R}}$ to denote the extended real line $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$. We use $\mathcal{L}_p(\Omega)$ for short whenever no ambiguity arises.

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

Convexity:

$\varrho(\gamma X + (1-\gamma)Y) \leq \gamma\varrho(X) + (1-\gamma)\varrho(Y)$

for all X, Y *and* $\gamma \in [0, 1]$.

Monotonicity:

If $X_\omega \geq Y_\omega$ for P-a.a $\omega \in \Omega$, then $\varrho(X) \geq \varrho(Y)$.

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 [94] 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 know as a dual representation of coherent measures of risk. 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

$$\langle \zeta, Z \rangle = \int_\Omega \zeta(\omega) Z(\omega) dP(\omega), \ \zeta \in \mathcal{L}_q(\Omega), \ Z \in \mathcal{L}_p(\Omega). \tag{2.6}$$

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 $dQ/dP = \zeta$.

Theorem 3 ([94])**.** *If ϱ is a finite-valued coherent measure of risk on $\mathcal{L}_p$, where $1 \leq p < \infty$, then 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*

$$\varrho(Z) = \sup_{\zeta \in \mathcal{A}} \langle \zeta, Z \rangle = \sup_{dQ/dP \in \mathcal{A}} \mathbb{E}_Q[Z]. \tag{2.7}$$

For every coherent measure of risk, the set $\mathcal{A}$ is the convex subdifferential of the functional $\varrho(\cdot)$ calculated at 0, i.e., $\mathcal{A} = \partial\varrho(0)$. We note that 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 $F_X(\eta) = P\{X \leq \eta\}$, we consider the survival function

$$\bar{F}_X(\eta) = P(X > \eta)$$

and the left-continuous inverse of the cumulative distribution function defined as follows:

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

It is clear that $F_X^{(-1)}(\alpha)$ is the left α-quantile of X.

We intend to apply the theory to investigate the distribution of classification errors and that is why we have a preference to small outcomes (small errors). We define the *Value at Risk* at level α of a random error X by setting

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

which implies that

$$P(V > \mathrm{VaR}_\alpha(X)) \leq \alpha.$$

The risk here is defined as the probability of the error X obtaining a large value. For a given α, we can minimize the value at risk by appropriately selecting the parameters of the classifier. This point of view corresponds to minimizing the probability of misclassification. Although Value at Risk is intuitively appealing measure, it is not coherent.

In the theory of measures of risk a special role is played by the functional called the Average Value-at-Risk and denoted $\mathrm{AVaR}(\cdot)$ (see [1, 76, 87]). The *Average Value at Risk* of X at level α is defined as

$$\mathrm{AVaR}_\alpha(X) = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_t(X)\,dt. \tag{2.8}$$

Consider the integrated survival function of the random variable X,

$$\bar{F}_X^{(2)}(\eta) = \int_\eta^\infty \bar{F}_X(t)\big]\, dt = \mathbb{E}[(X-\eta)_+].$$

The second equality is shown in [26]. The *upper Lorenz function* $\bar{F}_X^{(-2)} : \mathbb{R} \to \overline{\mathbb{R}}$ is introduced in [26] as a counterpart of the absolute Lorenz function (cf. [64, 2, 37]). It is defined as follows:

$$\bar{F}_X^{(-2)}(\alpha) = \int_\alpha^1 F_X^{-1}(t)\, dt \quad \text{for } 0 < \alpha < 1. \tag{2.9}$$

Additionally, $\bar{F}_X^{(-2)}(1) = 0$, $\bar{F}_X^{(-2)}(0) = \mathbb{E}(X)$, and $\bar{F}_X^{(-2)}(\alpha) = -\infty$ for $\alpha \notin [0,1]$. The function $\bar{F}_X^{(2)}(\cdot)$ is concave because its derivative is monotonically non-increasing.

Recall that, for a convex function $f : \mathbb{R}^n \to \mathbb{R}$, its Fenchel conjugate function, f^*, is defined as follows:

$$f^*(w) = \sup_v \{\langle v, w\rangle - f(v)\}.$$

The following result is shown in [26].

Theorem 4. *The Fenchel conjugate function of the integrated survival function $\bar{F}_X^{(2)}(\cdot)$ is the function $-\bar{F}_X^{(-2)}(\cdot + 1)$. Furthermore, whenever η is a α-quantile of X, where $\alpha \in (0,1)$, then*

$$\mathbb{E}(Z-\eta)_+ - \eta(\alpha - 1) = \bar{F}_X^{(-2)}(\alpha).$$

This statement is a counterpart of the conjugate duality relation for the absolute Lorenz curve, which has been first established in [74, Theorem 3.1]. From the definition of the upper Lorenz function, we obtain that it represents the Average Value-at-Risk:

$$\bar{F}_X^{(-2)}(1-\alpha) = \int_{1-\alpha}^1 \mathrm{VaR}_{1-t}(X)\, dt = \int_0^\alpha \mathrm{VaR}_\beta(X)\, d\beta \quad \text{for } 0 < \alpha < 1. \tag{2.10}$$

We obtain

$$\mathrm{AVaR}_\alpha(X) = \frac{1}{\alpha}\bar{F}_X^{(-2)}(1-\alpha) \quad \text{for } 0 < \alpha < 1.$$

Thus, using Theorem 4, we obtain

$$\begin{aligned}\mathrm{AVaR}_\alpha(X) &= \frac{1}{\alpha}\bar{F}_X^{(-2)}(1-\alpha) \\ &= -\frac{1}{\alpha}\sup_\eta\Big\{-\alpha\eta - \mathbb{E}[\max(0, X-\eta)]\Big\} \\ &= \inf_\eta\Big\{\eta + \frac{1}{\alpha}\mathbb{E}[\max(0, X-\eta)]\Big\}.\end{aligned}$$

This is the representation (cf. also [94]) suitable for optimization problems.

Due to Kusuoka theorem ([56],[94, 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 Average Value-at-Risk at all probability levels. This result can be extended for finite probability spaces with equally likely observations. Kusuoka representations allows to extend statistical estimators of Lorenz curves to spectral law-invariant measures of risk as shown in [27]. Central limit theorems for general composite risk functionals is established in [29].

Other popular coherent measures of risk (when small outcomes are preferred) include the upper mean-semi-deviations of order p, defined as

$$\sigma_p^+[Z] := \mathbb{E}[Z] + \kappa\Big(\mathbb{E}\Big[\big(Z - \mathbb{E}[Z]\big)_+^p\Big]\Big)^{1/p}, \tag{2.11}$$

where $p \in [1, \infty)$ is a fixed parameter. It is well defined for all random variables Z with finite p-th order moments and is coherent for $\kappa \in [0, 1]$. In the special case of $p = 1$, the upper semi-deviation is equal to $1/2$ of the absolute deviation, i.e.,

$$\mathbb{E}\Big[\big(Z - \mathbb{E}[Z]\big)_+\Big] = \frac{1}{2}\mathbb{E}\Big[\big|Z - \mathbb{E}[Z]\big|\Big]$$

Other classes of coherent measures of risk were proposed and analyzed in [20, 28, 55, 76, 94] and the references therein.

In [86], 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 the report [**?**]. 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. Our point of view is that *different attitude should be allowed to classification errors for the different classes.*

2.5 Risk Sharing Preliminaries

The notion of risk sharing and analysis of this topic 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. Risk allocation assumes that there is a quantitative assessment undertaken by a higher authority within a firm, which divides the firm's costs between the constituents. 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, which is defined as follows. Given convex functions $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots k$, their *infimal convolution* is the function $f_1 \square \cdots \square f_k : \mathbb{R}^n \to \mathbb{R}$ (see,[88, p. 57]) defined by

$$[f_1 \square \cdots \square f_k](x) = \inf\{f_1(x_1) + \cdots + f_k(x_k) \ : \ x_1 + \cdots + x_k = x\}.$$

The infimal convolution is a convex function and its Fenchel-conjugate satisfies is the

sum of the conjugate function f_i^*, $i = 1, \ldots, k$, i.e.,

$$[f_1 \square \cdots \square f_k]^* = f_1^* + \cdots + f_k^*.$$

The risk-sharing problem amounts to the evaluation of the infimal convolution

$$[\varrho_1 \square \cdots \square \varrho_k](X).$$

It is observed (see, e.g., [58, 65]) that the random variables X^i, $i = 1, \ldots, k$, which solve this problem, satisfy a co-monotonicity property as follows

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

We shall discuss the optimality of a risk allocation decision in due course. At the moment, we note that the problem setting and the results associate with risk sharing of losses 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.

2.6 Risk Sharing in Classification

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:

$$\begin{aligned} \min \; & \sum_{i=1}^{k} \mathbb{E}\big[Z^i(\pi)\big] \\ \text{subject to } \; & Z^i(\pi) = \operatorname{dist}(\varphi(X^i; \pi), K_i), \quad i = 1, \ldots, k, \\ & \pi \in \mathcal{D}. \end{aligned} \tag{2.12}$$

A formulation in line with the least-square approach in statistics uses the second-order moments as follows

$$\begin{aligned} \min \ & \sum_{i=1}^{k} \mathbb{E}\big[(Z^i(\pi))^2\big] \\ \text{subject to } \ & Z^i(\pi) = \operatorname{dist}(\varphi(X^i;\pi), K_i), \quad i = 1,\dots,k, \\ & \pi \in \mathcal{D}. \end{aligned} \tag{2.13}$$

We shall introduce the notion of a risk-averse classifier. Let a set of labeled data, a parametric classifier family $\varphi(\cdot;\pi)$ with the associated collection of sets K_i, $i = 1\dots,k$, and the law-invariant coherent risk measures ϱ_i, $i = 1\dots,k$ 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.

We assume throughout that the set of feasible parameters π is a closed convex set $\mathcal{D} \subseteq \mathcal{R}^s$. Let $\mathcal{Y}$ denote the set of all random vectors $(Z^1(\pi),\dots,Z^k(\pi))$ obtained as $Z^i(\pi) = \operatorname{dist}(\varphi(X^i;\pi), K_i)$, 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:

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

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

Definition 5. *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*

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

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 6. *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 extend-able to the case of more general orders defined by pointed cones.

Definition 7. *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 8. *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*

$$\mathbb{P}(Z > \eta) \geq \mathbb{P}(Z' > \eta) \quad \forall\, \eta \in \mathbb{R}, \tag{2.14}$$

or, equivalently, $F_Z(\eta) \leq F_{Z'}(\eta)$.

The relation is strict (denoted $Z \succ_{(1)} Z'$), if additionally, inequality (2.14) *is strict for some $\eta \in \mathbb{R}$.*

A random vector $\mathbf{Z} = (Z_1, \ldots Z_k)$ is stochastically larger than $\mathbf{Z}' = (Z'_1, \ldots Z'_k)$ (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., [92].

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

Theorem 9. *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 10. *If* ϱ *is a law invariant strictly monotonic coherent measure of risk, then*

$$\begin{aligned} &\varrho(Z) > 0 \text{ for all random variables } Z \geq 0 \text{ a.s.}, Z \not\equiv 0 \\ &\varrho(Z) < 0 \text{ for all random variables } Z \leq 0 \text{ a.s.}, Z \not\equiv 0. \end{aligned} \tag{2.15}$$

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 $\varrho_i(Z^i(\pi)) \geq 0$ for all $\pi \in \mathcal{D}$ and for all $i = 1, \dots k$ and, therefore, the attainable allocations lie in the positive orthant, i.e., $\mathcal{X} \subseteq \mathbb{R}^k_+$.

We assume everywhere that the risk measures ϱ^i used for evaluation of classification errors in classes $i = 1, ..., k$ are coherent, law invariant, and finite-valued.

Theorem 11. *Assume that the random vectors X^i, $i = 1, \dots k$, have bounded support. If the function $\varphi(x, \cdot)$ is continuous for every argument $x \in \mathbb{R}^n$ and the sets K_i, $i = 1, \dots k$ are non-empty, closed and convex, 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. The distance functions $z \mapsto \operatorname{dist}(z, K_i)$ are continuous convex functions (see, e.g., [4]) 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. Furthermore, the variables Z^i have bounded support due to the boundedness assumption of the theorem. Therefore, $Z^i(\cdot)$ is continuous with respect to the norm in the space $\mathcal{L}_p(\Omega)$. Since the risk measures $\varrho_i(\cdot)$ are convex and finite, they are continuous on $\mathcal{L}_p$ for $p \geq 1$. 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

$$\|\varphi(x;\pi) - z^i(\pi)\| = \min_{z\in K_i} \|\varphi(x;\pi) - z\| \tag{2.16}$$

$$\|\varphi(x;\pi) - z^i(\pi')\| = \min_{z\in K_i} \|\varphi(x;\pi') - z\| \tag{2.17}$$

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,

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

The monotonicity and convexity axioms for the risk measures imply that

$$\varrho_i\big(\operatorname{dist}(\varphi(X;\pi_\lambda), K_i)\big) \le \lambda\varrho_i\big(\operatorname{dist}(\varphi(X;\pi), K_i)\big) + (1-\lambda)\varrho_i\big(\operatorname{dist}(\varphi(X;\pi'), K_i)\big).$$

□

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 12. *Assume that the function $\varphi(x,\cdot)$ is affine for every argument $x \in \mathbb{R}^n$, the sets $\mathcal{D} \subseteq \mathbb{R}^s$, and K_i $i = 1,\dots k$ are non-empty, closed, and convex. 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*

$$\min_{\pi \in \mathcal{D}} \sum_{i=1}^{k} w_i \varrho_i\big(\operatorname{dist}(\varphi(X_i;\pi), K_i)\big). \tag{2.18}$$

Proof. Statement follows form the well-known scalarization theorem in vector optimization problems ([69]) and Theorem 11. □

Theorem 13. *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 8.*

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 $\varrho_i(Z^i(\pi)) \geq \varrho_i(Z^i(\pi'))$ 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 14. *Suppose each component of the vector function* $\varphi(x,\cdot)$ *is affine for every* $x \in \mathbb{R}^n$ *and the sets* $\mathcal{D}$ *and* K_i*,* $i = 1, \dots, k$*, are non-empty, convex, and closed. 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*

$$\min_{\pi \in \mathcal{D}} \sum_{i=1}^{k} \sum_{j=1}^{m_i} \mu_j^i \operatorname{dist}(\varphi(x_j^i;\pi), K_i). \tag{2.19}$$

Proof. Since the parameter $\hat{\pi}$ defines a Pareto-optimal classifier $\varphi(\cdot, \hat{\pi})$ for the RSCP and all conditions of Corollary 12 are satisfied, then $\hat{\pi}$ is an optimal solution of problem (2.18) 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 (2.7). 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 (2.7) is achieved at some elements $\zeta^i \in \mathcal{A}_i$. We have $\zeta^i_j \geq 0$, $\sum_{j=1}^{m_i} \frac{\zeta^i_j}{m_i} = 1$ because ζ_i are probability densities. We obtain

$$\varrho_i(\mathrm{dist}(\varphi(X^i;\pi), K_i)) = \sum_{j=1}^{m_i} \frac{\zeta^i_j}{m_i} \mathrm{dist}(\varphi(x^i_j;\pi), K_i).$$

Setting

$$\mu^i_j = w_i \frac{\zeta^i_j}{m_i},\ j = 1, \ldots, m_i,\ i = 1, \ldots, k$$

we observe that the vector $\mu \in R^{m_1 + \ldots m_k}$ constitutes a probability mass function. Thus, problem (2.18) can be reformulated as (2.19). □

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, $i = 1, \ldots, k$.

2.7 Optimization of Risk Sharing

We analyze the risk-sharing classification problem (2.18) with the purpose of suggesting a way of treating it numerically efficiently. First, we formulate optimality conditions for this problem. The composite nature of the problem (2.18) is difficult and that is why we reformulate the problem. We introduce auxiliary variables $Y \in \mathcal{L}_p(\Omega, \mathcal{F}, P; \mathbb{R}^m)$, $i = 1, \ldots k$, which are defined by the constraints:

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

Problem (2.18) can be reformulated to

$$\begin{aligned} &\min_{\pi,Y} \sum_{i=1}^{k} w_i \varrho_i(\|Y^i\|) \\ &\text{s.t. } \varphi(X^i;\pi) + Y^i \in K_i, \quad \forall i = 1,\dots,k, \\ &\qquad \pi \in \mathcal{D}. \end{aligned} \tag{2.20}$$

We shall show that this problem is equivalent to (2.18).

Lemma 15. *For any solution $\hat{\pi}$ of problem* (2.18)*, random vectors $\hat{Y}^i$ exist, so that $(\hat{\pi}, \hat{Y})$ solves problem* (2.20) *as well, where $\hat{Y} = (\hat{Y}^k, \dots, \hat{Y}^k)$ and for any solution $(\hat{\pi}, \hat{Y})$ of problem* (2.20)*, the vector $\hat{\pi}$ is a solution of problem* (2.18) *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:

$$Y^i(\omega) = \mathrm{Proj}_{K_i}\big((\varphi(X(\omega);\pi)\big) - \varphi(X(\omega);\pi). \tag{2.21}$$

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 equation (2.21). □

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 (2.20) 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 (2.20).

We denote the realizations of the random vectors Y^i, $i = 1,\dots,k$, 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 (2.20):

$$\begin{aligned} &\min_{\pi,Y} \ \sum_{i=1}^{k} w_i \varrho_i(\|Y^i\|) \\ &\text{s.t.} \ \ \varphi(x_j^i;\pi) + y_j^i \in K_i, \quad \forall j = 1,\dots,m_i,\ i = 1,\dots,k, \\ &\qquad \pi \in \mathcal{D}. \end{aligned} \tag{2.22}$$

Theorem 16. *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 (2.22) if and only if probability mass functions $\zeta^i \in \partial\varrho_i(0)$ and vectors g_j^i from $\partial\|\hat y_j^i\|$ exist such that*

$$0 \in -\sum_{i=1}^{k}\sum_{j=1}^{m_i} w_i \zeta_j^i (g_j^i)^\top D\varphi(X^i;\hat\pi) + \mathcal{N}_{\mathcal{D}}(\hat\pi) \tag{2.23}$$

$$w_i \zeta_j^i g_j^i \in \mathcal{N}_i\big(\varphi(x_j^i;\hat\pi) + \hat y_j^i\big) \ \textit{for all} \ \ j = 1,\dots m_i,\ i = 1,\dots k. \tag{2.24}$$

Proof. We assign Lagrange multipliers λ_j^i to the inclusion constraints and define the Lagrange function as follows:

$$L(\pi, Y, \lambda) = \sum_{i=1}^{k} \Big(w_i \varrho_i(\|Y^i\|) + \sum_{j=1}^{m_i} \big\langle \varphi(x_j^i;\pi) + y_j^i, \lambda_j^i \big\rangle \Big).$$

Using optimality conditions [9, Theorem 3.4], we obtain that $(\hat\pi, \hat Y)$ is optimal for problem (2.22) if and only if $\hat\lambda$ exists such that

$$\begin{aligned} 0 &\in \partial_{(\pi,Y)} L(\hat\pi, \hat Y, \hat\lambda) + \mathcal{N}(\hat\pi, \hat Y) \\ \hat\lambda_j^i &\in \mathcal{N}_i\big(\varphi(x_j^i;\hat\pi) + \hat y_j^i\big). \end{aligned}$$

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

$$0 \in \sum_{i=1}^{k} \sum_{j=1}^{m_i} (\hat{\lambda}_j^i)^\top D\varphi(x_j^i; \hat{\pi}) + \mathcal{N}_{\mathcal{D}}(\hat{\pi}) \tag{2.25}$$

$$0 = w_i \partial_Y \varrho_i(\|Y\|) + \hat{\lambda}^i, \quad i = 1, \dots k, \tag{2.26}$$

$$\hat{\lambda}_j^i \in \mathcal{N}_i\big(\varphi(x_j^i; \hat{\pi}) + \hat{y}_j^i\big), \quad j = 1, \dots, m_i,\ i = 1, \dots k. \tag{2.27}$$

We calculate the multipliers $\hat{\lambda}^i$ from the equation (2.26) 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 (2.25) and (2.27), we obtain condition (2.23) and (2.24). □

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 (2.22) can be reformulated as a risk-averse two-stage optimization problem (cf. [93]). The first stage decision is π and the first stage problem is

$$\min_{\pi \in \mathcal{D}} \sum_{i=1}^{k} w_i \varrho_i\big(Z^i(\pi)\big)\big). \tag{2.28}$$

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

$$z_j^i(\pi) = \min_{y \in K_i} \|\varphi(x_j^i; \pi) - y\|, \quad j = 1, \dots m_i,\ i = 1, \dots k. \tag{2.29}$$

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 (2.29) can be computed as follows:

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

Then the optimal value of (2.29) is

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

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

2.8 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 [29]. In [29], a composite risk functional is defined in the following way.

$$\varrho(X) = \mathbb{E}\left[f_1\left(\mathbb{E}\left[f_2\left(\mathbb{E}\left[\cdots f_\ell\left(\mathbb{E}\left[f_{\ell+1}\left(X\right)\right], X\right)\right] \cdots, X\right)\right], X\right)\right] \tag{2.30}$$

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

$$\begin{aligned} \varrho^{(N)} = \sum_{i_0=1}^{N} \frac{1}{N} \Big[f_1 \Big(\sum_{i_1=1}^{N} \frac{1}{N} \big[f_2 \big(\sum_{i_2=1}^{N} \frac{1}{N} \big[\cdots f_\ell \big(\sum_{i_\ell=1}^{N} \frac{1}{N} f_{\ell+1}(X_{i_\ell}), X_{i_{\ell-1}} \big) \big] \\ \cdots, X_{i_1} \big) \big], X_{i_0} \Big) \Big] \end{aligned} \tag{2.31}$$

It is shown in [29] that the most popular measures of risk fit the structure (2.30). 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 $f_j\big(\cdot, x\big)$ at points μ_{j+1} in directions ζ_{j+1}. It is given by

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

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 [29]. Furthermore, it is shown that similar construction represents the Average-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 (2.30) by choosing the innermost function $f_{\ell+1}(x) : \mathbb{R}^d \to \mathbb{R}$ to be $f_{\ell+1}(x) = \mathrm{dist}(\varphi(x;\pi), K_i)$ whenever φ satisfies properties i–iii.

In our setting each misclassification risk $\varrho_i\Big(\mathrm{dist}\,\big(\varphi(X^i;\pi), K_i\big)\Big)$ is estimated by $\varrho_i^{(m_i)}\big(\|\hat{Y}^i\|\big)$, where $(\hat{Y}^i; \hat{\pi})$ is the solution of problem (2.22). 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:

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

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 (2.31). 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 ϱ is given by

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

2.9 Risk Sharing in SVM

We analyze the SVM problem in more detail. We consider only law-invariant 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-allocation $\mathcal{X}$ for the affine classifier $\varphi(z;\pi) = \langle v, z\rangle - \gamma$. Due to Corollary‘12, we can determine a risk-averse classifier by solving the following problem:

$$\begin{aligned}
\min_{v,\gamma,Z^1,Z^2}\ & \lambda\varrho_1(Z^1) + (1-\lambda)\varrho_2(Z^2) \\
\text{s. t. }\ & \langle v, x_j^1\rangle - \gamma + z_j^1 \geq 0, \quad j = 1,\dots,m_1, \\
& \langle v, x_j^2\rangle - \gamma - z_j^2 \leq 0, \quad j = 1,\dots,m_2, \\
& \langle v, v\rangle = 1, \\
& Z^1 \geq 0,\ Z^2 \geq 0.
\end{aligned} \tag{2.32}$$

Here $\lambda \in (0,1)$ is a parameter and the vectors Z^i have realization z_j^i, $i = 1,2$ and $j = 1,\dots,m_i$, representing the classification error for the sample of each class. The

random vectors Z^i can be represented by a deterministic vectors stacking all realizations z_j^i as components (sub-vectors) of it. Abusing notation, we shall use Z^i also for those long vectors in $\mathbb{R}^{nm_i}$.

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

$$\gamma_M = \max_{1\leq j\leq m_i,\, i=1,2} v^\top x_j^i \quad \gamma_m = \min_{1\leq j\leq m_i,\, i=1,2} v^\top x_j^i.$$

Thus, in this case, we can set $\mathcal{D} = \mathbb{R}^n$. We also consider a soft-margin risk-averse SVM based on problem (2.3), although the classification error might not be calculated properly. The problem reads

$$\begin{aligned} \min_{v,\gamma,Z^1,Z^2} \;& \lambda\varrho_1(Z^1) + (1-\lambda)\varrho_2(Z^2) + \delta\|v\|^2 \\ \text{s. t.} \;& \langle v, x_j^1\rangle - \gamma + z_j^1 \geq 1, \quad j = 1,\ldots,m_1, \\ & \langle v, x_j^2\rangle - \gamma - z_j^2 \leq -1, \quad j = 1,\ldots,m_2, \\ & Z^1 \geq 0,\, Z^2 \geq 0. \end{aligned} \tag{2.33}$$

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 (2.32).

We observe that the parameter (v, γ) for each Pareto-optimal classifier can be obtained by solving the following problem:

$$\begin{aligned} \min_{v,\gamma,Z^1,Z^2} \;& \varrho_1(Z^1) + \varrho_2(Z^2) \\ \text{s. t.} \;& \langle v, x_i^1\rangle - \gamma + \frac{1}{\lambda} z_i^1 \geq 0, \quad i = 1,\ldots,m_1, \\ & \langle v, x_j^2\rangle - \gamma - \frac{1}{1-\lambda} z_j^2 \leq 0, \quad j = 1,\ldots,m_2, \\ & \langle v, v\rangle = 1, \\ & Z^1 \geq 0,\, Z^2 \geq 0. \end{aligned} \tag{2.34}$$

Lemma 17. *Problem* (2.34) *is equivalent to problem* (2.32).

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. □

This observation is a counterpart of the result in [48] for the risk sharing of random losses among constituents.

In order to solve problem (2.32) numerically, we use sequential local convex approximation to problem (2.32). 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:

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

The following auxiliary problem is a convex approximation of problem (2.32):

$$\begin{aligned} \min_{v,\gamma,Z^1,Z^2} \quad & \lambda\varrho_1(Z^1) + (1-\lambda)\varrho_2(Z^2) \\ \text{s. t.} \quad & \langle v, x_i^1\rangle - \gamma + z_i^1 \geq 0, \quad i = 1,\ldots,m_1, \\ & \langle v, x_j^2\rangle - \gamma - z_j^2 \leq 0, \quad j = 1,\ldots,m_2, \\ & \langle \bar{v}, v\rangle = 1, \\ & Z^1 \geq 0,\ Z^2 \geq 0. \end{aligned} \tag{2.35}$$

We shall denote the objective function by f:

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

Observe that for a vector v whose norm is different than one, 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 (2.32).

Step 0. Set $\ell = 1$; choose initial points v_1 and γ_1 with $\|v_1\| = 1$ and calculate the corresponding Z_1^1, Z_1^2, and $f(Z_1^1, Z_1^2, v_1, \gamma_1)$.

Step 1. Solve problem (2.35) with $\bar{v} = v_\ell$.

Denote its solution by $(Z_{\ell+1}^1, Z_{\ell+1}^2, \hat{v}_\ell, \gamma_{\ell+1})$.

Step 2. If $f(Z_{\ell+1}^1, Z_{\ell+1}^2, \hat{v}_\ell, \gamma_{\ell+1}) = f(Z_\ell^1, Z_\ell^2, v_\ell, \gamma_\ell)$, then stop; otherwise set $v_{\ell+1} = \frac{\hat{v}_\ell}{\|\hat{v}_\ell\|}$, increase ℓ by one and go to Step 1.

When the method stops, then the point $(Z_\ell^1, Z_\ell^2, v_\ell, \gamma_\ell)$ satisfies the optimality conditions for problem (2.32). Otherwise, the method generates a sequence of points $\{(Z_\ell^1, Z_\ell^2, v_\ell, \gamma_\ell)\}_{\ell=1}^\infty$, which converges to a point satisfying the optimality conditions for problem (2.32).

2.10 Kernel-based Risk-averse Binary Classification

We adopt the formulation (2.33) and use Average-Value at Risk at level $\alpha \in (0,1)$ for both classes. Assume that we have chosen a kernel K with associated mapping

ψ. The classification problem becomes

$$\min_{v,\gamma,t_1,t_2,Z^1,Z^2,Y^1,Y^2} \quad \lambda\big(t_1 + \frac{1}{\alpha m_1}\sum_{j=1}^{m_1} y_j^1\big) + (1-\lambda)\big(t_2 + \frac{1}{\alpha m_2}\sum_{j=1}^{m_2} y_j^2\big) + \delta\|v\|^2 \quad (2.36)$$

$$\text{s. t.} \quad \langle v, \psi(x_j^1)\rangle - \gamma + z_j^1 \geq 1, \quad j = 1, \ldots, m_1, \quad (2.37)$$

$$\langle v, \psi(x_j^2)\rangle - \gamma - z_j^2 \leq -1, \quad j = 1, \ldots, m_2, \quad (2.38)$$

$$y_j^i \geq z_j^i - t_i, \quad j = 1, \ldots, m_i,\ i = 1, 2, \quad (2.39)$$

$$Z^1 \geq 0,\ Z^2 \geq 0,\ Y^1 \geq 0,\ Y^2 \geq 0. \quad (2.40)$$

Without loss of generality, we may assume $\gamma = 0$.

We assign Lagrange multipliers μ_j^1, $j = 1, \ldots, m_1$, to constraints (2.37) and μ_j^2, $j = 1, \ldots, m_2$, to constraints (2.38), respectively. The Lagrange multipliers associated with constraints (2.39) are denoted ζ_j^i, $j = 1, \ldots, m_i,\ i = 1, 2$. The Lagrange function has the form

$$\begin{aligned} L(v, \gamma, t_1, t_2, Z^1, Z^2, Y^1, Y^2, \mu^1, \mu^2, \zeta^1, \zeta^2) = \\ \lambda\big(t_1 + \frac{1}{\alpha m_1}\sum_{j=1}^{m_1} y_j^1\big) + (1-\lambda)\big(t_2 + \frac{1}{\alpha m_2}\sum_{j=1}^{m_2} y_j^2\big) + \delta\|v\|^2 \\ + \sum_{j=1}^{m_1} \mu_j^1(1 - \langle v, \psi(x_j^1)\rangle - z_j^1) + \sum_{j=1}^{m_2} \mu_j^2(1 + \langle v, \psi(x_j^2)\rangle - z_j^2) \\ + \sum_{j=1}^{m_1} \zeta_j^1(z_j^1 - t_1 - y_j^1) + \sum_{j=1}^{m_2} \zeta_j^2(z_j^2 - t_2 - y_j^2). \end{aligned} \quad (2.41)$$

The optimality conditions for problem (2.36)–(2.40) yield

$$2\delta v = \sum_{j=1}^{m_1} \mu_j^1 \psi(x_j^1) - \sum_{j=1}^{m_2} \mu_j^2 \psi(x_j^2) \tag{2.42}$$

$$\lambda = \sum_{j=1}^{m_1} \zeta_j^1 \tag{2.43}$$

$$1 - \lambda = \sum_{j=1}^{m_2} \zeta_j^2 \tag{2.44}$$

$$0 \le \zeta_j^1 \le \frac{\lambda}{\alpha m_1}, \quad j = 1, \dots m_1 \tag{2.45}$$

$$0 \le \zeta_j^2 \le \frac{1-\lambda}{\alpha m_2}, \quad j = 1, \dots m_2 \tag{2.46}$$

$$0 \le \mu_j^i \le \zeta_j^i \quad j = 1, \dots m_i, \ i = 1, 2. \tag{2.47}$$

$$\tag{2.48}$$

Let $\hat{v}$ be the optimal solution when minimizing the Lagrangian with respect to the non-negativity constraints (2.40). Using the optimality conditions, the dual function becomes

$$\delta \|\hat{v}\|^2 - \sum_{j=1}^{m_1} \mu_j^1 \langle \hat{v}, \psi(x_j^1) \rangle + \sum_{j=1}^{m_2} \mu_j^2 \langle \hat{v}, \psi(x_j^2) \rangle + \sum_{j=1}^{m_1} \mu_j^1 + \sum_{j=1}^{m_2} \mu_j^2 \tag{2.49}$$

$$= \delta \|\hat{v}\|^2 - \langle \hat{v}, \sum_{j=1}^{m_1} \mu_j^1 \psi(x_j^1) - \sum_{j=1}^{m_2} \mu_j^2 \psi(x_j^2) \rangle + \sum_{j=1}^{m_1} \mu_j^1 + \sum_{j=1}^{m_2} \mu_j^2 \tag{2.50}$$

$$= \delta \|\hat{v}\|^2 - 2\delta \langle \hat{v}, \hat{v} \rangle = -\delta \|\hat{v}\|^2 + \sum_{j=1}^{m_1} \mu_j^1 + \sum_{j=1}^{m_2} \mu_j^2. \tag{2.51}$$

Substituting the form of $\hat{v}$ from the optimality conditions, we obtain the following form of the dual function:

$$-\frac{1}{4\delta} \Big[\sum_{i=1}^{2} \sum_{j=1}^{m_i} \sum_{\ell=1}^{m_i} \mu_j^i \mu_\ell^i \langle \psi(x_j^i), \psi(x_\ell^i) \rangle - 2 \sum_{j=1}^{m_1} \sum_{\ell}^{m_2} \mu_j^1 \mu_\ell^2 \langle \psi(x_j^1), \psi(x_\ell^2) \rangle \Big] + \sum_{j=1}^{m_1} \mu_j^1 + \sum_{j=1}^{m_2} \mu_j^2.$$

Using the form of the kernel, the dual function becomes

$$-\frac{1}{4\delta} \Big[\sum_{i=1}^{2} \sum_{j=1}^{m_i} \sum_{\ell=1}^{m_i} \mu_j^i \mu_\ell^i K(x_j^i, x_j^i) - 2 \sum_{j=1}^{m_1} \sum_{\ell=1}^{m_2} \mu_j^1 \mu_\ell^2 K(x_j^1, x_\ell^2) \Big] + \sum_{j=1}^{m_1} \mu_j^1 + \sum_{j=1}^{m_2} \mu_j^2.$$

This form is valid under the relations given by the optimality conditions. We eliminate the ζ variables by noticing that they induce the constraints

$$\begin{aligned} 0 \leq \mu_j^1 &\leq \frac{\lambda}{\alpha m_1} \\ 0 \leq \mu_j^2 &\leq \frac{1-\lambda}{\alpha m_2} \\ \sum_{j=1}^{m_1} \mu_j^1 &\leq \sum_{j=1}^{m_1} \zeta_j^1 = \lambda \\ \sum_{j=1}^{m_2} \mu_j^2 &\leq \sum_{j=1}^{m_2} \zeta_j^2 = 1-\lambda \end{aligned}$$

Using the form of the kernel, the dual problem to problem (2.36)–(2.40) can be formulated as follows:

$$\begin{aligned} \min \ & \sum_{i=1}^{2} \sum_{j=1}^{m_i} \sum_{\ell=1}^{m_i} \mu_j^i \mu_\ell^i K(x_j^i, x_\ell^i) - 2 \sum_{j=1}^{m_1} \sum_{\ell=1}^{m_2} \mu_j^1 \mu_\ell^2 K(x_j^1, x_\ell^2) - 4\delta \Big[\sum_{j=1}^{m_1} \mu_j^1 + \sum_{j=1}^{m_2} \mu_j^2 \Big] \\ \text{s. t.} \quad & \sum_{j=1}^{m_1} \mu_j^1 \leq \lambda, \\ & \sum_{j=1}^{m_2} \mu_j^2 \leq 1-\lambda \\ & 0 \leq \mu_j^1 \leq \frac{\lambda}{\alpha m_1}, \quad j = 1, \dots m_1, \\ & 0 \leq \mu_j^2 \leq \frac{1-\lambda}{\alpha m_2}, \quad j = 1, \dots m_2. \end{aligned}$$

After solving the dual problem, we obtain $\hat{\mu}^i, j,\ j = 1, \dots, m_i,\ i = 1, 2$. To calculate the value of the classifier for a new observation x, we calculate

$$\begin{aligned} \langle v, x \rangle &= \frac{1}{2\delta} \sum_{j=1}^{m_1} \hat{\mu}_j^1 \langle \psi(x_j^1), \psi(x) \rangle - \frac{1}{2\delta} \sum_{j=1}^{m_2} \hat{\mu}_j^2 \langle \psi(x_j^2), \psi(x) \rangle \\ &= \frac{1}{2\delta} \sum_{j=1}^{m_1} \hat{\mu}_j^1 K(x_j^1, x) - \frac{1}{2\delta} \sum_{j=1}^{m_2} \hat{\mu}_j^2 K(x_j^2, x). \end{aligned}$$

Therefore, we do not need to know the form of the mapping ψ and the kernel trick works for the RSSVM as well.

2.11 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 F_1–score. Further, we showcase the flexibility of the framework by exploring the Pareto-efficient frontier of various classifiers derived from our framework. In our numerical experiments, we have used the Average Value-at-Risk and the mean semi-deviation of order one.

2.11.1 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 [60]. 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 the following table.

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

Table 2.1: Data summary

2.11.2 Model Formulations

We consider several scenarios for choices of measures of risk. In the first scenario, we treat one of the classes (Class0) in a risk neutral manner, while applying the mean-semi-deviation measure to the classification error of the second class. We call this loss function "asym_risk" (see Table 2.2). 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 Average Value-at-Risk of the classification error. These convex combinations use an additional model parameter $\beta \in (0,1)$. We note that such a convex combination is a coherent measure of risk. Recall that AVaR_α can be calculated by a linear optimization problem.

$$\mathrm{AVaR}_\alpha(Z) = \min_\eta \Big\{ \eta + \frac{1}{\alpha}\mathbb{E}\big[Z-\eta\big]_+ \Big\}$$

The formulation (2.33) for these loss function require modification due to the use of the variational form of the Average-Value at Risk at level $\alpha \in (0,1)$. We have already formulated the problem for minimizing the Average-Value at Risk at level $\alpha \in (0,1)$ for both classes in section 2.10 in the case of using a mapping to a higher dimensional space.

Table 2.2 displays the chosen combinations of risk measure pairs for the binary classification scenario in order to give an easy overview.

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

Table 2.2: Risk measure combinations used as loss functions in the experiments

We note that calculation of the first order semi-deviation and the average 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. Further experiments are necessary to access the effect of higher order measures of risk.

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 (2.3). The second risk-neutral formulation uses the Huber loss function and leads to the following problem formulation

$$\begin{aligned} \min_{v,\gamma,Z^1,Z^2} \;& \frac{1}{m_1}\sum_{i=1}^{m_1} \min\left(z_i^1, (z_i^1)^2\right) + \frac{1}{m_2}\sum_{j=1}^{m_2} \min\left(z_j^2, (z_j^2)^2\right) + \delta\|v\|^2 \\ \text{s. t. } & \langle v, x_i^1\rangle - \gamma + z_i^1 \geq 1, \quad i = 1,\ldots,m_1, \\ & \langle v, x_j^2\rangle - \gamma - z_j^2 \leq -1, \quad j = 1,\ldots,m_2, \\ & Z^1 \geq 0,\, Z^2 \geq 0. \end{aligned} \tag{2.52}$$

The third benchmark uses a single risk measure (2.53) on the total error as proposed in [**?**]. It has the following formulation.

$$\begin{aligned} \min_{v,\gamma,t,Z^1,Z^2,Y^1,Y^2} \;& \beta\Big(\frac{1}{m_1}\sum_{j=1}^{m_1} z_j^1 + \frac{1}{m_2}\sum_{j=1}^{m_2} z_j^2\Big) + \\ & \qquad (1-\beta)\bigg(t + \frac{1}{\alpha(m_1+m_2)}\Big(\sum_{j=1}^{m_1} y_j^1 + \sum_{j=1}^{m_2} y_j^2\Big)\bigg) + \delta\|v\|^2 \\ \text{s. t. } & \langle v, x_i^1\rangle - \gamma + z_i^1 \geq 1, \quad i = 1,\ldots,m_1, \\ & \langle v, x_j^2\rangle - \gamma - z_j^2 \leq -1, \quad j = 1,\ldots,m_2, \\ & y_j^i \geq z_j^i - t, \quad j = 1,\ldots,m_i,\; i = 1,2, \\ & Z^1 \geq 0,\, Z^2 \geq 0,\, Y^1 \geq 0,\, Y^2 \geq 0. \end{aligned} \tag{2.53}$$

Interestingly, both risk-neutral formulations produce 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 Average Value-at-Risk measure, as "joint_cvar".

The problem formulations which we use in our experiments are the following.

Expected value vs. Average Value-at-Risk – "asym_risk"

$$\begin{aligned}
\min_{v,\gamma,t,Z^1,Z^2,Y} \quad & \frac{\lambda}{m_1}\sum_{j=1}^{m_1} z_j^1 + \frac{1-\lambda}{m_2}\sum_{j=1}^{m_2}(y_j + z_j^2) + \delta\|v\|^2 \\
\text{s. t.} \quad & \langle v, x_j^1\rangle - \gamma + z_j^1 \geq 1, \quad j = 1,\dots,m_1, \\
& \langle v, x_j^2\rangle - \gamma - z_j^2 \leq -1, \quad j = 1,\dots,m_2, \\
& y_j \geq z_j^2 - t, \quad j = 1,\dots,m_2, \\
& Z^1 \geq 0,\ Z^2 \geq 0,\ Y \geq 0.
\end{aligned} \tag{2.54}$$

Mean-semi-deviation vs. Average Value-at-Risk – "one_cvar"

$$\begin{aligned}
\min_{v,\gamma,t,Z^1,Z^2,Y^1,Y^2} \quad & \frac{\lambda}{m_1}\sum_{j=1}^{m_1}(y_j^1 + z_j^1) + (1-\lambda)\big(t + \frac{1}{\alpha m_2}\sum_{j=1}^{m_2} y_j^2\big) + \delta\|v\|^2 \\
\text{s. t.} \quad & \langle v, x_j^1\rangle - \gamma + z_j^1 \geq 1, \quad j = 1,\dots,m_1, \\
& \langle v, x_j^2\rangle - \gamma - z_j^2 \leq -1, \quad j = 1,\dots,m_2, \\
& y_j^1 \geq z_j^1 - \frac{1}{m_1}\sum_{j=1}^{m_1} z_j^1, \quad j = 1,\dots,m_1, \\
& y_j^2 \geq z_j^2 - t, \quad j = 1,\dots,m_2, \\
& Z^1 \geq 0,\ Z^2 \geq 0,\ Y^1 \geq 0,\ Y^2 \geq 0.
\end{aligned} \tag{2.55}$$

Mean-semi-deviation vs. Expecatation and AVaR – "risk_cvar"

$$\begin{aligned}
\min_{v,\gamma,t,Z^1,Z^2,Y^1,Y^2} \quad & \frac{\lambda}{m_1}\sum_{j=1}^{m_1}(y_j^1 + z_j^1) + \frac{\beta(1-\lambda)}{m_1}\sum_{j=1}^{m_2} z_j^2 \\
& \qquad + (1-\beta)(1-\lambda)\left(t + \frac{1}{\alpha m_2}\sum_{j=1}^{m_2} y_j^2\right) + \delta\|v\|^2 \\
\text{s. t.} \quad & \langle v, x_j^1\rangle - \gamma + z_j^1 \geq 1, \quad j = 1,\dots,m_1, \\
& \langle v, x_j^2\rangle - \gamma - z_j^2 \leq -1, \quad j = 1,\dots,m_2, \\
& y_j^1 \geq z_j^1 - \frac{1}{m_1}\sum_{j=1}^{m_1} z_j^1, \quad j = 1,\dots,m_1, \\
& y_j^2 \geq z_j^2 - t, \quad j = 1,\dots,m_2, \\
& Z^1 \geq 0,\ Z^2 \geq 0,\ Y^1 \geq 0,\ Y^2 \geq 0.
\end{aligned} \tag{2.56}$$

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

$$\begin{aligned}
\min_{v,\gamma,Z^1,Z^2,Y^1,Y^2} \quad & \frac{\lambda}{m_1}\sum_{j=1}^{m_1}(y_j^1+z_j^1)+\frac{1-\lambda}{m_2}\sum_{j=1}^{m_2}(y_j^2+z_j^2)+\delta\|v\|^2 \\
\text{s. t.} \quad & \langle v, x_j^1\rangle-\gamma+z_j^1\geq 1, \quad j=1,\ldots,m_1, \\
& \langle v, x_j^2\rangle-\gamma-z_j^2\leq -1, \quad j=1,\ldots,m_2, \\
& y_j^i\geq z_j^i-\frac{1}{m_i}\sum_{j=1}^{m_i} z_j^i,, \quad j=1,\ldots,m_i,\ i=1,2, \\
& Z^1\geq 0,\ Z^2\geq 0,\ Y^1\geq 0,\ Y^2\geq 0.
\end{aligned} \tag{2.57}$$

Average-Value at Risk for both classes – "two_cvar"

$$\begin{aligned}
\min_{v,\gamma,t_1,t_2,Z^1,Z^2,Y^1,Y^2} \quad & \delta\|v\|^2+\lambda\beta_1\sum_{j=1}^{m_1} z_j^1+\lambda(1-\beta_1)\left(t_1+\frac{1}{\alpha m_1}\sum_{j=1}^{m_1} y_j^1\right) \\
& +(1-\lambda)\beta_2\sum_{j=1}^{m_1} z_j^2+(1-\lambda)(1-\beta_2)\left(t_2+\frac{1}{\alpha m_2}\sum_{j=1}^{m_2} y_j^2\right) \\
\text{s. t.} \quad & \langle v, x_j^1\rangle-\gamma+z_j^1\geq 1, \quad j=1,\ldots,m_1, \\
& \langle v, x_j^2\rangle-\gamma-z_j^2\leq -1, \quad j=1,\ldots,m_2, \\
& y_j^i\geq z_j^i-t_i, \quad j=1,\ldots,m_i,\ i=1,2, \\
& Z^1\geq 0,\ Z^2\geq 0,\ Y^1\geq 0,\ Y^2\geq 0.
\end{aligned} \tag{2.58}$$

2.11.3 Performance

We perform k-fold cross-validation and all reported results are out of sample. In Tables 2.3, 2.5, and 2.7, we report the F_1–score and AUC, along with recall, 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 one of results results optimized for each metric. We use this highlight the additional flexibility that the proposed method introduces, in the next section.

F_1-score Optimized Classifiers

	exp_val	joint_cvar	asym_risk	one_cvar	risk_cvar	two_risk	two_cvar
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

	exp_val	joint_cvar	asym_risk	one_cvar	risk_cvar	two_risk	two_cvar
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**

Table 2.3: Main results table for the WDBC dataset – Displaying the model parameters for the each model formulation as well as the corresponding performance metrics.

In the above Table 2.3, 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 formuations.

Further, this formulation does have the best value for the competing metric in

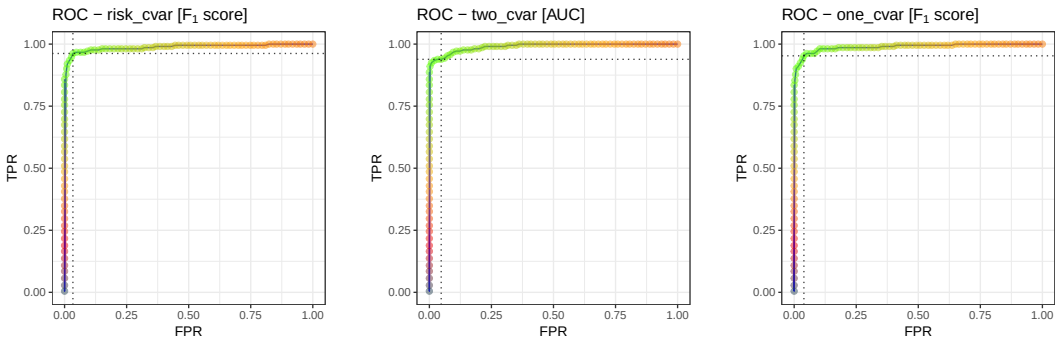

Figure 2.2: ROC plots for the best performing model formulations on the WDBC data: "risk_cvar" with the best F_1-score, "two_cvar" with the best AUC value, and "one_cvar" for the alternate metric.

both cases. The respective ROC curves for each of the classifiers are displayed in Figure 2.2. 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 2.4 contains the calulations of risk, with respect to each model formulation. More specifically, for each obtained classifier we caculate 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 avarage value at risk for the α quantiles 0.75, 0.85, and 0.95.

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 performes

WDBC

		Expectation	MSD	$\text{AVaR}_{0.75}$	$\text{AVaR}_{0.85}$	$\text{AVaR}_{0.95}$
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

Table 2.4: Risk Evalutation for the WDBC data set – Displaying the expectation of error, Mean Semi-deviation, and Avarage Value at Risk for the α quantiles 0.75, 0.85, and 0.95

best with respect to the AUC metric is the worst performing, benchmarks excluded. Looking closely at the corresponding ROC curve in Figure 2.2 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 2.5 we observe that the best performing model with respect to F_1-score is the again "risk_cvar" model with 0.68581 compared to the 0.66785 of the risk neutral formulations. Similarly, the "one_cvar" model is again second in this conext, 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 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.

F_1-score Optimized Classifiers

	exp_val	joint_cvar	asym_risk	one_cvar	risk_cvar	two_risk	two_cvar
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

	exp_val	joint_cvar	asym_risk	one_cvar	risk_cvar	two_risk	two_cvar
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

Table 2.5: Main results table for the "pima-indians-diabetes" dataset – Displaying the model parameters for the each model formulation as well as the corresponding performance metrics.

Looking closely at the ROC curves in Figure 2.3 we can see that the AUC pioritized "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.

Figure 2.4 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

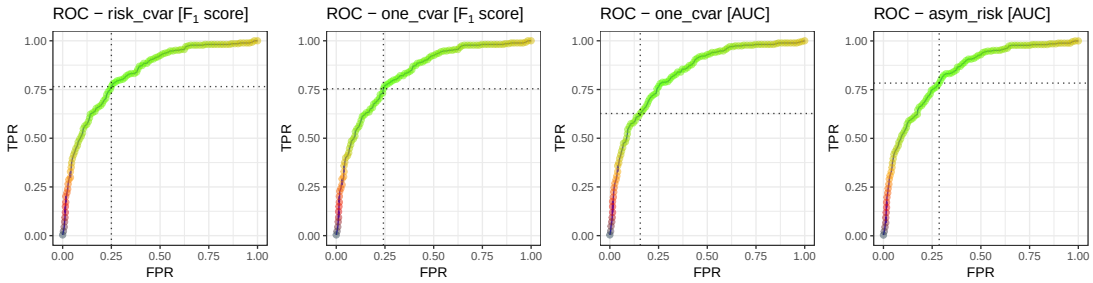

Figure 2.3: ROC plots for the best performing model formulations on the "pima-indians-diabetes" data: "risk_cvar" with the best F_1-score, "one_cvar" featuring both parameter sets, and finally the "asym_risk" formulation featuring the best AUC value

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 eachother and the benchmarks. Virtually no distinction can be made between the ROC curves for the various classifiers. However, looking at the error distribution plot on the left, we notice that the the two benchmarks misclassify less of the default class and more of the target class. On the other hand, the "two_cvar" formulation underperforms for the opposite reason, in relation to the target metric and the best performing formulation "risk_cvar".

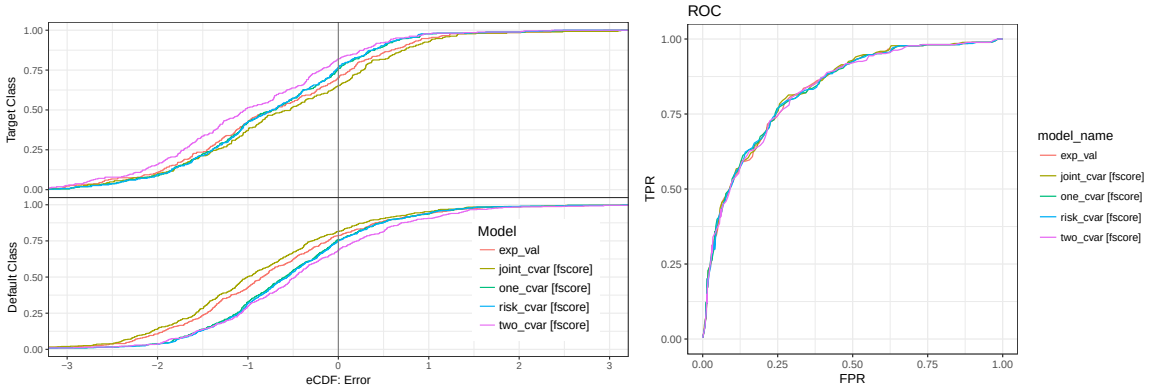

Figure 2.4: Empirical distribution of error realizations comparing risk-averse loss function formulations to benchmarks [F_1-score] on the "pima-indians-diabetes" dataset (left) and the corresponding ROC curves (right)

pima-indians-diabetes

		Expectation	MSD	$\text{AVaR}_{0.75}$	$\text{AVaR}_{0.85}$	$\text{AVaR}_{0.95}$
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

Table 2.6: Risk Evalutation for the "pima-indians-diabetes" data set – Displaying the expectation of error, Mean Semi-deviation, and Avarage Value at Risk for the α quantiles 0.75, 0.85, and 0.95

Table 2.7 contains the risk functional evalutation 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 evalution on the third and final dataset, whose main performance metrics are shown in Table 2.7. 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 varibles.

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"

F_1-score Optimized Classifiers

	exp_val	joint_cvar	asym_risk	one_cvar	risk_cvar	two_risk	two_cvar
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

	exp_val	joint_cvar	asym_risk	one_cvar	risk_cvar	two_risk	two_cvar
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**

Table 2.7: Main results table for the "seismic-bumps" dataset – Displaying the model parameters for the each model formulation as well as the corresponding performance metrics.

formulation that leads group, but again at a significant cost of the F_1-score. Looking at Figure 2.5, we can see room for improvemnts to the this by changing the threshold on the AUC prioritzed "two_cvar" model. We observe that in terms of stability to that respect, the "asy_risk" formulation along with "joint_cvar" benchmark have less variation.

Turning the attention to the risk functional evaluation in Table 2.8, 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.

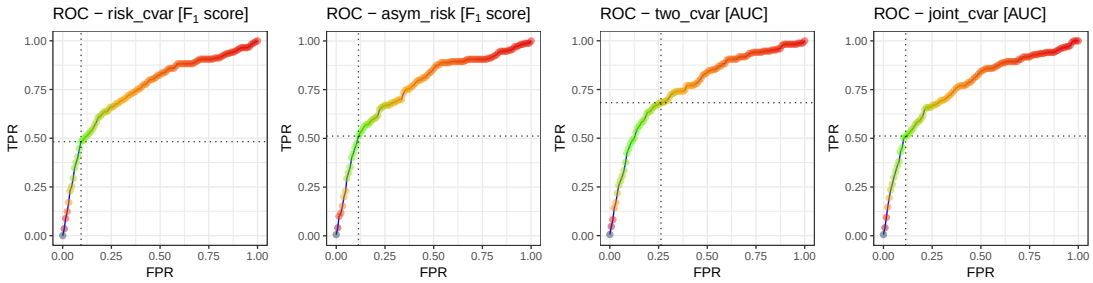

Figure 2.5: ROC plots for the best performing model formulations on the "seismic-bumps" data: "risk_cvar" with the best F_1-score, "one_cvar", "joint_cvar", "two_cvar" formulation featuring the best AUC value

Notice, in Figure 2.6, 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.

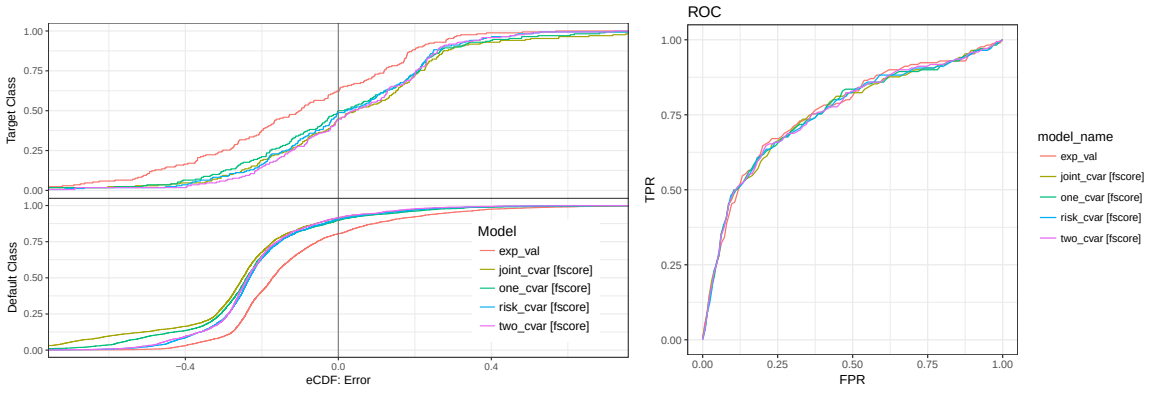

Figure 2.6: Empirical distribution of error realizations comparing risk-averse loss function formulations to benchmarks [F_1-score] on the "seismic-bumps" dataset (left) and the corresponding ROC curves (right)

seismic-bumps

		Expectation	MSD	$AVaR_{0.75}$	$AVaR_{0.85}$	$AVaR_{0.95}$
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

Table 2.8: Risk Evalutation for the "seismic-bumps" data set – Displaying the expectation of error, Mean Semi-deviation, and Avarage Value at Risk for the α quantiles 0.75, 0.85, and 0.95

2.11.4 Flexibility

Our approach provides additional flexibility which is generally not available for classification methods like SVM. We allow the user to implement a predetermined attitude toward risk of misclassification, and to explore the Pareto-efficient frontier of classifiers. The efficient frontier can be used to chose a risk-averse classifier according additional criterion as the F_1–score, AUC, or other similar performance metrics, as discussed in the previsou section.

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 Figures 2.7 , 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

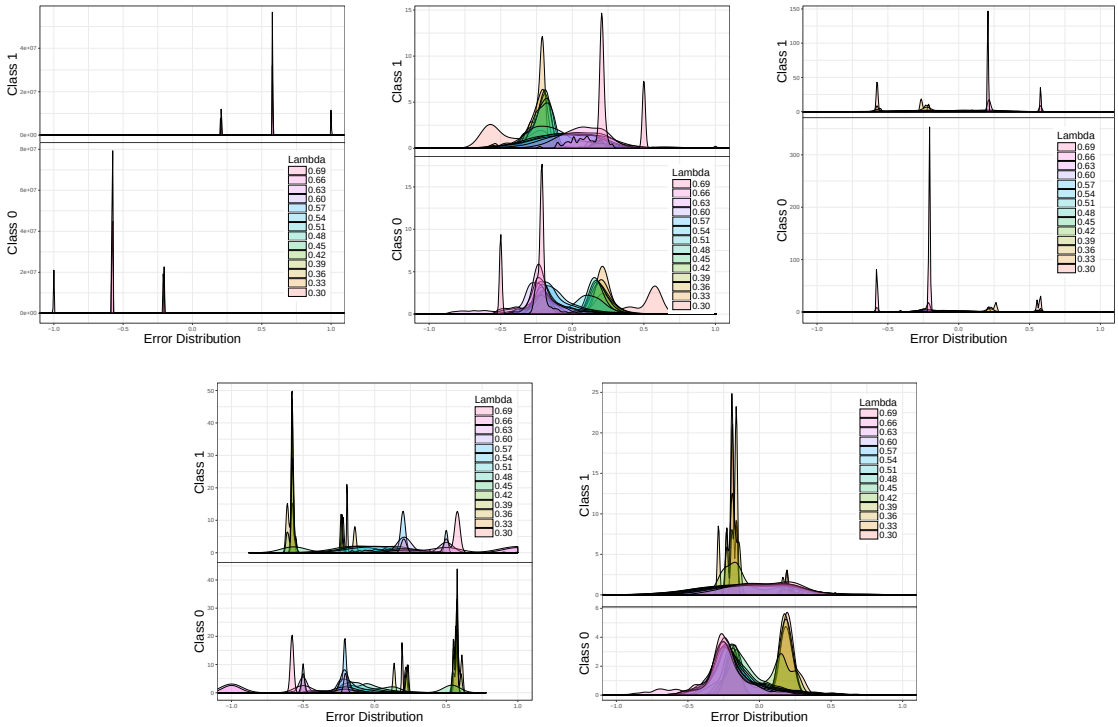

Figure 2.7: The distribution of error displayed as smoothed histogram for each of five proposed formulations for the risk-averse SVM problem e.g. "asym_risk", "one_cvar", "risk_cvar", "two_risk", and "two_cvar" all using the same set of λ values, with other parameters fixed, on the "seismic-bumps" dataset

of risk-averse Pareto-optimal classifiers.

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 Average Value-at-Risk in a similar way. We observe that the loss function "one_cvar" consistently provides the best performance. 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.

2.12 Concluding remarks

This thesis proposes a novel approach to classification problems by leveraging mathematical models of risk. We have formulated several optimization problems for optimizing a classifier over a parametric family of functions. The problem's objective is a weighted sum of risk measures, associated with the classification error of the classes: each class may be treated with an individual risk preference. We have shown the existence of an optimal risk-sharing classifier under mild assumptions. Additionally, we have demonstrated that the optimal risk-sharing classifier also solves a risk-neutral classification problem, in which the empirical probabilities of the data points are replaced by a probability distribution determined implicitly by the risk measures. The risk-averse classification problem provides a robust classifier due to the dual representation of risk measures.

We have provided a more specific problem formulation for the case of binary classification and have shown how the methodology allows for the use of kernel functions. Additionally, we have proposed an efficient numerical method for solving a version of the binary risk-averse classification by determining a separating plane with a normal vector of length one. This allows for precise calculation of the classification error.

We have conducted experiments on three data sets and we have compared our approach to three benchmarks, which use the minimization of the total expected error, the Huber function, and the Average Value-at-Risk for the total classification error as presented in [**?**]. Our observations are the following. On the data sets for which traditional formulations perform well, the novel approach performs on par or slightly better depending on the particular choice of risk measures and parameters. The proposed approach has an advantage on all data sets as measured by the F_1-score. Exploring the Pareto-efficient frontier provides additional flexibility and is a tool for customizing the classifier. As we see from the numerical results, we achieve larger recall or precision by adjusting the scalarization factor λ. Overall, this is

an extremely flexible approach which allows fine-tuning leading allowing the user to achieve the best possible result in the chosen metric.

Chapter 3

The Impact of Patent Activity on Stock Dynamics in the Technology Sector

3.1 Introduction

The prospect of a company in the stock market is traditionally evaluated by financial statement analysis [23]. However, the performances of high-tech companies on the stock market are also strongly related to their technology innovations, which are often protected by patents. Also, while stock forecasting is a topic of general interest and has been studied extensively in the literature, much less efforts have been devoted to the study of the impact of patent activities on stock movement patterns. Indeed, rich information about the prospects of the high-tech companies is available in patent data. For instance, patent data can be used to reveal the characteristics of innovative activities in the companies, such as emerging technologies developed in the companies, knowledge diffusion and technological change in the companies, and the global strategies of firms. These characteristics are important factors of long-term future market performances of companies. Here, we assume that research and development investments, strategic technological investments, and market aspirations of a company are reflected in its patent activity. Also, we consider the quantity of developed patented innovations in a given period as well as the span of those innovations in terms of patent category relevant factors for the company's development. Indeed, in this thesis, we focus on the analysis and discovery of the relations between patent activities of a company and its market performance. Our aim is to provide a new perspective and to illustrate the potential of exploiting the information available in

patent data for stock prospecting.

Along these lines, we investigate the relationship between the patent activities of high-tech companies and the return of the company's stock, its drift and volatility. We adopt the most popular model for the movement of a stock price over time and use it to determine the drift and volatility changes over a relevant time horizon. The drift of a stock refers to the changes of the average return. Volatility in stock markets refers to the variance range of stock tracking. We emphasize that, while the drift of a stock is essential, the volatility is crucial both as a risk factor, as well as an indicator of investment opportunities. In the stock market, volatility is a major indicator of market stability. Volatility within a certain range indicates stock market running in a steady pace. However, if the stock price exhibits high volatility, from a market macro-structure point of view, the risk of the whole financial system will be increased. From a micro-structure point of view, the investors will face a higher risk on their investment. Proper evaluation of the company's prospects may aid the decision whether a certain investment should be increased or decreased in a period of high volatility. Additionally, the internal review of a company's efficiency is often judged by the market performance of its stock. Therefore, it is very important to monitor and perhaps even anticipate large fluctuations in stock prices. In the literature, researchers investigate stock volatility directly through the trading data. This usually provides a short-term view of the market stability and investment prospects. In contrast, the analysis of the impact of patent activities on stock volatility has the promise to provide a long-term view of the market stability of a high-tech company.

To the best of our knowledge, we are the first to attempt to marry the areas of patent data mining and financial modeling. In the context of this investigation, we have developed a new approach to establish relations between the patent activities of high-tech companies and the dynamics of their stock price movement. Specifically, we develop a stochastic model to characterize the relationships between the monthly drift/volatility of stock prices and patent activities, which have been reflected by the

number of patent applications and the diversity of the corresponding patent categories. To gain an insight into the patent activities, we extract all patent events in the form of applications as well as their categories for every company we have decided to investigate. Also, we determine how many patents contribute to new innovation categories and how many new categories are produced each month. For the financial data, we employ daily stock trading data to calculate returns. We consider the influence of the market and adjust the return to extract the movement of the stock price, which is due purely to the activity and the market evaluation of the specific company and not to the general movements of the entire market. Additionally, by considering various lagged terms, we produce a number of fitted models with a moving window technique. We then analyze how the average performance and the volatility relate to the chosen factors (patent activity indicators). Finally, we perform statistical testing on the models and their coefficients to establish statistical significance.

The validation has been performed on real-world stock trading data as well as patent data. We notice that there is a statistically significant impact of the patent activities on the market-adjusted stock return process, as well as on its drift and volatility. While known approaches to relating sequential data consists of relating the time series directly, our new method reveals the impact of the chosen patent-activity indicators on the stock's volatility, which would be impossible to infer otherwise. The results confirm the impact of innovations on stock price movement and show that the stock prices can exhibit more volatility if the company has been extending their patents to new areas. This is reflected in the positive coefficients for innovation terms in our models for volatility.

3.2 A Literature Review

3.2.1 Stochastic Models of Dynamic Systems

We mention here the most popular statistical model for time-series data.

Autoregressive Moving Average (ARMA) models allow us to describe stochastic processes in terms of two polynomials by combining a purely auto-regressive model and a moving average model. A critical assumption, for all three model types, is that the process being described be weakly stationary, and since strict stationarity does not exist in practice, from here on out referred to as stationary. Given a time series X_t, AR(p) model has the following form

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \epsilon_t. \tag{3.1}$$

Model MA(q) has the following form

$$X_t = \mu + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \tag{3.2}$$

where $\mu = \mathbb{E}[X_t]$, $\theta_1, ..., \theta_p$'s are model parameters and $\epsilon_t, \epsilon_{t-1}, ..., \epsilon_{t-p}$ are error terms.

Combining (3.1) and (3.2) we obtain the form for the ARMA(p, q) model

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \tag{3.3}$$

The models are used under the assumption that the error terms $\epsilon_t = \{\epsilon_1, ...\epsilon_T\}$ are independent identically distributed random variables following a normal distribution with mean zero and equal variance. This assumption is frequently used to test model fit.

Generalized Autoregressive Conditional Heteroskedastic (GARCH) models, first introduced by Bollerslev in [7], are a generalization of the ARCH model introduced by Engle in [33]. Both use an exact function to govern the evolution of the conditional standard deviation of a time series, commonly referred to as **volatility**.

Let F_t be a filtration on X_t, which is the time series we wish to model. We can denote the conditional mean and variance of X_t given F_{t-1} as follows

$$\begin{aligned} \mu_t &= \mathbb{E}(X_t|F_{t-1}) \\ \sigma_t^2 &= \mathrm{Var}(X_t|F_{t-1}) = \mathbb{E}[(X_t - \mu_t)^2|F_{t-1}] \end{aligned} \tag{3.4}$$

Further, suppose, X_t follows some simple time series model, a stationary ARMA(p, q) for instance. That is, we can consider the following model structure for X_t

$$\begin{aligned} X_t &= \mu_t + a_t \\ \mu_t &= \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{i=1}^{q} \theta_i a_{t-i} \\ Y_t &= X_t - \phi_0 - \sum_{i=1}^{k} \beta_i r_{it} \end{aligned} \tag{3.5}$$

where k,p and q are nonnegative integers, and r_{it} are the explanatory variables associated with the ARMA(p, q) model. We use Y_t to denote the time series after adjusting for the effect of the mean equation μ_t. Combining (3.4) and (3.5), we obtain

$$\sigma_t^2 = \mathrm{Var}(X_t|F_{t-1}) = \mathrm{Var}(a_t|F_{t-1}) \tag{3.6}$$

In many econometric and finance problems one needs to model the amount of increase or decrease of investments per time period. For that purpose, ARCH models are created. They model the changing variance of a time series. The ARCH(m) Model has the form

$$\begin{aligned} a_t &= \sigma_t \epsilon_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_m a_{t-m}^2 \end{aligned} \tag{3.7}$$

where ϵ_t is a sequence of independent identically distributed random variables with mean zero and variance 1, $\alpha_0 > 0$ and $\alpha_i \geq 0$ for $i > 0$ A more general model is a GARCH (generalized autoregressive conditionally heteroscedastic) model. It uses the past squared observations and past variances to model the variance at a given time.

GARCH(m, s) Model has the form

$$\begin{aligned} a_t &= \sigma_t \epsilon_t \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2 \end{aligned} \tag{3.8}$$

where ϵ_t is a sequence of iid random variables with mean zero and variance 1, $\alpha_0 > 0$ and $\alpha_i \geq 0$, $\beta_j \geq 0$ and $\sum_{i=1}^{\max(m,s)}(\alpha_i + \beta_i) < 1$. Here it is understood that $\alpha_i = 0$ for $i > m$ and $\beta_j = 0$ for $j > s$.

In practice, it is often assumed that ϵ_t follow a standard normal, standardized Student-t distribution, or a generalized error distribution.

Markov chains are the most popular models of stochastic processes. We assume that the process, which we observe takes finitely many states, which we number, i.e., the state space is $\mathcal{S} = \{1, \ldots, S\}$. Markov chain is a good model for such a system, if we determine that only the current state is necessary for the identification of the probability distribution of the next state of the system. Mathematically, the following property is required. A random process X is called a Markov chain if it has the following Markov property

$$\mathbb{P}(X_n = i | X_0 = x_0, X_1 = x_1, \cdots, X_{n-1} = x_{n-1}) = \mathbb{P}(X_n = i | X_{n-1} = x_{n-1})$$

for all $n \geq 1$ and all states $i, x_0, x_1, \ldots, x_{n-1} \in S$.

The Markov chain X is called homogeneous, if

$$\mathbb{P}(X_{n+1} = j | X_n = i) = \mathbb{P}(X_1 = j | X_0 = i).$$

The probabilities $p_{ij} = \mathbb{P}(X_n = j | X_{n-1} = i)$ are called **transition probabilities**. This chain is completely characterized by its transition probabilities comprised in the matrix $P = (p_{ij})_{i,j \in S}$, which is called the transition matrix. The n-step transition matrix $P_n = (p_{ij}(n))_{i,j \in S}$ is the matrix of the n-step transition probabilities:

$$p_{ij}(n) = \mathbb{P}(X_{n+m} = j | X_m = i) = \mathbb{P}(X_n = j | X_0 = i).$$

It is well-known that $P_n = P^n$. The unconditional probability distribution of any state of the system X_n can be obtained from the initial probability distribution and the transition matrix. Denoting the mass function by a row vector $\pi^{(n)}$, i.e., $\pi_i^{(n)} = \mathbb{P}(X_n = i)$, we have $\pi^{(m+n)} = \pi^{(m)} P_n$, and, hence, $\pi^{(n)} = \pi^{(0)} P^n$ with $\pi^{(0)}$ being the initial distribution. The distribution π is called stationary distribution of the chain if

$$\pi = \pi P, \quad \text{i.}e., \quad \pi_j = \sum_{i \in S} \pi_i p_{ij} \text{ for all } j \in S.$$

The assumption of stationarity is implicit in data mining applications, in particular in Bayesian learning models.

One example of how we can associate a Markov chain with a collection of time-series is the following. We assume that each time-series is given by vector in $\mathbb{R}^n$ and all time-series constitute the set $\mathcal{S} \subset \mathbb{R}^n$. Let ϱ be a similarity measure which we shall use to compare the points of $\mathcal{S}$. We assume that in addition to the axioms of the definition ϱ is positive semi-definite. This means, that for all bounded functions $f : \mathcal{S} \to \mathbb{R}$, we have

$$\sum_{x,y \in \mathcal{S}} \varrho(x, y) f(x) f(y) \geq 0.$$

We define the following general relation of a time-series $x \in \mathcal{S}$:

$$d(x) = \sum_{y \in \mathcal{S}} \varrho(x, y). \tag{3.9}$$

Then, we can associate a Markov transition matrix P defined by setting

$$p_{x,y} = \frac{\varrho(x, y)}{d(x)}.$$

We can easily verify that P is a stochastic matrix because its entries are non-negative and the elements in each row sum to 1. These are the probabilities to move from state x to another state in one time step, i.e., the conditional distribution of the future state given the current state x. These probabilities provide a normalized measure of similarity between the time-series in the data-base. The stationary distribution of

the Markov chain with this transition matrix is proportional to the total association $d(x)$.

$$\pi(y) = \frac{d(y)}{\sum_{z \in \mathcal{S}} d(z)}.$$

Geometric Brownian Motion Models are use to model many physical processes. Recently, this model became popular also for processes in the area of finance. Originally, the Brownian motion was introduce to describe the motion of particles suspended in a fluid. The idea is that the motion of the particle is due to the sum of a large number of very small random forces.

A stochastic process $\{W_t; t0\}$ with continuous sample paths is called standard Brownian motion if it has the following properties.

1. $W_0 = 0$.

2. W has independent stationary increments $W_t - W_s$, $0 \leq s < t$.

3. $W_t - W_s$ has a normal distribution with mean zero and variance $t - s$, $0 \leq s < t$.

A process X_t is a Brownian motion with variance σ and drift μ if it has the form

$$X_t = \sigma W_t + \mu t.$$

A Brownian motion with drift has properties 1. and 2. above but property 3. is modified to state that the distribution of $X_t X_s$ has a normal distribution with mean (ts) and variance σ.

Geometric Brownian motion can be defined as the process $\{S_t; t0\}$ defined by

$$S_t = S_0 e^{X_t},$$

where X_t is a Brownian motion with variance σ and drift μ and $S_0 > 0$ is the initial value. Using this formula, it can be shown that the Geometric Brownian Motion satisfies the Markov Chain property. This type of stochastic processes are adopted in quantitative finance for the purpose of describing the dynamics of stock prices. We shall use it in our analysis as well.

3.2.2 Patent Data Analysis

There is significant amount of research into leveraging patent data in a variety of ways. Many researchers have attempted to parse the actual text of various parts of patent documents [98, 39, 63] in order to ascribe economic value to a patent, or a set of patents, or alternatively to establish quality metrics for ranking patents. For instance, the authors [98] introduced "SIMPLE," a piece of software designed to provide an interactive way to analyze unstructured data such as patents and other scientific texts. This is accomplished by taking a set of keywords as input and generating multiple queries for each of them. Some of the keywords may be related and the software attempts to discover the relationship and structure the queries accordingly. The queries are then executed on the document database and the software provides detailed statistics about their occurrences within documents. Also, Hasan et al. [39] propose a method for discovering and ranking novel patents. The software starts by directly parsing the text in the "claims" section of the patent document, then rates patents based on how recent they are, and how impactful the phrases discovered in the text are. In addition, Liu et al. developed a latent graphical model in [63], which infers patent quality. They utilized natural language processing techniques in order to capture quality measures such as originality, clarity of claims, and importance of the prior works cited in the patent.

Another branch of research into patent data mining considers using topic models. For example, Tang [99] presented a system for mining patent data which uses topic models in order to facilitate the collection and analysis of the information from a heterogeneous patent network. Using a probabilistic model, they derived a ranking method for a set of patents, as well as the methods to summarize search results, and an efficient algorithm for the topic-level competitor evolution analysis.

Attempting to establish patent trends in order to gain competitive advantage in business is the focus in [97], where the authors develop an approach for identifying patent trends. They have considered company-level trends and related them to

industry-level trends by leveraging association rules. In contrast, the authors of [47] tried to minimize patent maintenance cost by proposing a method for analysis, prediction, and recommendation of patent maintenance policy. Here, the patents are modeled as a dynamic, heterogeneous information network, which changes with time.

Kim et al. adopted a citation network perspective in [52], and proposed a technique for automating the detection of influential patents via their novel centrality measure based on the change of a node similarity matrix. The authors alleviated the problem of computational intensity for centrality measures through new and clever way of updating values in the similarity matrix.

Unlike the above mentioned works, the focus of our thesis is on establishing relations between the patent activities of high-tech companies and the dynamics of their stock price movement. The goal is to show the promises of exploiting patent data for the analysis and prospecting of high-tech companies in the stock market.

In the literature, one of the most popular and widely accepted model of stock prices, respectively stock return, is based on the geometric Brownian motion process. In this setting, the stock price return is a solution of a stochastic differential equation. The essential ingredients of the model are a drift term and a volatility term. Substantial part of the research effort, in the quantitative finance community, is devoted to investigating the effect of various factors influencing the stock returns. To the best of our knowledge, all models relate the direct observations of the stock price or stock return to the value of the factors at a given time. For example, in [62], the author explored the efficiency impacts of Nikkei 225 future contracts on the underlying stocks. In [89], the predictive ability of economic indicators in the context of RTS index was studied. Also, in [81], the predictability of spot rates of the US dollar against British pound was investigated and the effect of brand acquisitions and disposals on the stock market was studied in [105].

The work presented in [79] and [72] are closely related to ours. Looking at patent

citations, the authors proposed an empirical strategy to estimate competition of companies based on their innovation record. In [79], the authors related the company's market return to information about their patent citations. In their paper, citation patterns are created and related to the area of science, in which the company patents. The authors have provided empirical evidence that markets positively reward companies when patents are granted and furthermore, the market value of the company increases when its patent portfolio is cited. In [72], the authors analyzed the effect of innovation markets on the market value of publicly traded companies in Japan. They developed a strategy using patent citation patterns to measures both patent importance and the emergence of potentially competing technologies. They suggested an estimator of market value, which addresses the potential endogeneity of R&D to company value.

Complementary to [79], we propose to investigate the impact of patent applications by looking separately at the innovations of the company in their traditional area of business and those which are in areas that are new to the company. Furthermore, we shall relate not only the stock return process but also its main characteristics, drift and volatility. Additionally, our basic assumptions are that a major factor explaining the movement of the stock price is the market movement itself. This is why we consider it important to account for this factor before analyzing the impact of patent activity. We explain the mathematical tools employed and further elaborate on their novelty in Section 3.3.

While our focus is primarily from the investment perspective, there is a significant amount of work that focuses on mining patent data for a myriad of other applications. To the best of our knowledge, there is no other work in this direction. Furthermore, our approach in establishing the impact of patent activity on the characteristics of the stock return process is novel and has not been proposed elsewhere.

3.3 The Adjusted Return Process

Undoubtedly, the overall stock market performance has a significant impact on individual stocks. This is the reason that we look at the returns of the stock only after removing that effect. This step allows us to address more accurately the impact of patent activities on the movement of the stock return. We call the obtained market-independent component of the stock return the *market-adjusted return*, whose calculation constitutes the first step of our data analysis. In a second step, we relate the return process to the selected patent activity factors. Subsequently, we develop a model to discover the impact of patent activity on the drift and volatility of the market-adjusted stock return process.

Basic models of stock prices assume that the price deviates from a certain steady state as a result of the trading process, i.e., ask and bid in financial markets. If we consider a stock with price S_t at time t and an expected rate of return μ, then the return or the relative change of the price during the next period of time dt is composed of two parts:

1. A part, which can be described as predictable, deterministic and anticipated, that is the expected return from the stock hold during a period of time dt; this return is assumed to be equal to $\mu S_t dt$.

2. A stochastic part, which reflects the random changes in stock prices during the interval of time dt attributed to external effects such as news, reports, random fluctuations in the market demands, etc. A frequently adopted assumption is that this contribution is proportional to the stock price. For a constant σ_t and a random walk process dW_t, this part of the return is assumed equal to $\sigma S_t \, dW_t$.

These modeling assumptions lead to the stochastic differential equation followed by the stock price (see, e.g., [71, Section 5.1]):

$$dS_t = \mu S_t \, dt + \sigma S_t \, dW_t \tag{3.10}$$

or, equivalently,

$$\frac{dS_t}{S_t} = \mu\, dt + \sigma\, dW_t \tag{3.11}$$

The stochastic differential Equation (3.10) describes the Brownian motion with drift followed by the stock price S_t. The return on S_t in the period of time dt follows an Itô's process. For every interval of time of length dt between two consecutive instants, the return can be represented as follows:

$$\frac{dS_t}{S_t} = d\big(\ln(S_t)\big) = \ln(S_t) - \ln(S_{t-dt}) = \ln\Big(\frac{S_t}{S_{t-dt}}\Big)$$

Thus, the model can be written as follows:

$$\ln\Big(\frac{S_t}{S_{t-dt}}\Big) = \mu\, dt + \sigma\, dW_t. \tag{3.12}$$

The solution of this stochastic differential equation is

$$\ln(S_t) = \ln(S_{t-dt}) + \big(\mu - \frac{1}{2}\sigma^2\big)\, dt + \sigma\varepsilon\sqrt{dt}' \tag{3.13}$$

(see e.g. [53, Chap. 4, p. 105]), or equivalently,

$$S_t = S_{t-dt} \exp\Big(\big(\mu - \frac{1}{2}\sigma^2\big)\, dt + \sigma\varepsilon\Big). \tag{3.14}$$

For the purpose of removing the market effect, we use the Capital Asset Pricing Model (CAPM), which involves one of the most significant characteristics of a stock, called Beta (β). In 1990, William Sharpe won a Nobel Prize in Economics for his work in developing the model [95]. In finance, the Beta of a stock or a portfolio is a number describing how the return of an asset is predicted by a benchmark. The Beta is usually estimated via the use of representative indices, such as the Standard and Poor 500 (S&P 500) index. We will discuss this issue in due course. It is assumed that Beta measures systematic risk based on how returns co-move with the overall market. Beta is also referred to as financial elasticity or correlated relative volatility. It has the following interpretation:

$\beta < 0$ Asset generally moves in the opposite direction as compared to the benchmark.

$\beta = 0$ Movement of the asset is uncorrelated with the movement of the benchmark.

$\beta \in (0, 1)$ Movement of the asset is generally in the same direction as, but less than the movement of the benchmark

$\beta = 1$ Movement of the asset is about the same as the movement of the benchmark

$\beta > 1$ Movement of the asset is generally in the same direction as, but more than the movement of the benchmark

Another theory, called the arbitrage pricing theory (APT) introduces multiple factors into consideration at the same time. According to this theory, the model has multiple betas associated with multiple risk factors. Each risk factor has a corresponding beta indicating the responsiveness of the asset being priced to that risk factor. Multiple-factor models contradict CAPM by claiming that some other factors can influence return significantly, therefore one may find two stocks with equal beta but they may not be equally good investment.

We adopt the point of view that the market is a major factor in the movement of the price with much grater significance than other possible factors. Therefore, our proposed approach accounts first for the market impact before other factors' influence on the price movement are investigated. We first adjust the returns (3.12) by removing the component related to the market.

More formally, let us consider the stochastic process of a representative index I_t over a period of time $t = 1 \dots T$. The CAPM model implies that the stock price S_t is related to the index I_t as follows:

$$\frac{dS_t}{S_t} - J_f = \beta(\frac{dI_t}{I_t} - J_f) + R_t. \tag{3.15}$$

Here J_f is the risk-free rate and R_t is the component of the process S_t, which represents the part of change not explained by the changes of the market as a whole.

The process R_t will be the subject of our investigation and we shall investigate how this process is influenced by the patent activity of a company in the technology sector. It can also be viewed as the premium over an investment in Exchange-traded fund (ETF) representing the index I_t. We shall refer to it the market-adjusted stock premium.

3.4 Patent-Activity Factors Affecting the Stock Return

In order to relate the adjusted return to the patent related indicators, we consult the multi-factor models establishing the impact of certain indicator of the financial instruments. Many studies are available on multi-factor models in asset pricing (see, e.g., [103, 96, 73, 8, 16, 19, 68]). Factors that change in time are of particular interest to risk premium forecasting and fit our goal as well. One of the multi-factor models most widely used in research and in practice is the APT model described in [19] as follows:

$$R_t = \alpha + \beta_1 I_t^1 + \beta_2 I_t^2 + \ldots \beta_k I_t^k, \tag{3.16}$$

where R_t stands for asset return and the factors I^j are major external economic factors, such as industrial production, inflation, interest rates, business cycle, etc. Models of this type relate the securities prices to the economic conditions. The same model is also applied to investment portfolios helping investors to determine factor sensitivity of their entire portfolio rather than of individual securities.

In [73], the authors also investigated a multi-factor model in which they separated the impacting quantities as static and dynamic. Another approach is suggested in [68], where the authors developed a time-series multi-factor portfolio analysis model, which was named as Constrained Flexible Least Squares (CFLS).

Our approach, which will be presented in Section 3.4.2, differs from the factor models in the extant literature. Due to the latent character of research and development with respect to stock market performance, we propose to extend model (3.16)

by introducing l lagged variables of the patent activity factors.

$$\begin{aligned} R_t = \alpha + \beta_0^1 I_t^1 + \beta_1^1 I_{t-1}^1 + ... + \beta_l^1 I_{t-l}^1 + \\ \beta_0^2 I_t^2 + \beta_1^2 I_{t-1}^2 + ... + \beta_l^2 I_{t-l}^2 + \\ \beta_0^3 I_t^3 + \beta_1^3 I_{t-1}^3 + ... + \beta_l^3 I_{t-l}^3 \end{aligned} \tag{3.17}$$

Note that we will use this extended version of the model as a baseline in our experiments. We consider this setting a new way of formulating factor models.

3.4.1 Consistency of Time Scales

For the process of returns R_t, we assume that a model analogous to (3.12) is valid. Also, we make the following modeling assumptions.

1. The process R_t is approximated by an aggregate process $\mathcal{R}_\tau$ of a larger time scale, where one time unit τ aggregates κ time intervals of the finer time-scale, for some natural number κ.

2. Within each time period $\tau = 1, \dots \Theta$ in the new time scale, the process R_t follows the following model

$$\begin{aligned} R_t = \mu_\tau \, dt + \sigma_\tau \, dW_t \\ t = \kappa\tau, \kappa\tau + 1, ..., \kappa * (\tau + 1) - 1, \end{aligned} \tag{3.18}$$

   with a constant drift μ_τ, and volatility σ_τ.

3. Patent activities of the company impact the drift and/or the volatility of the adjusted aggregate process $\mathcal{R}_\tau$.

The uniformity of aggregation is not essential for our analysis. Our methods apply equally well to a non-uniform time-scale, i.e., earlier periods may be aggregated in larger batches. In a non-homogeneous time frame, we need only set suitable time intervals $\tau = 1, \dots \Theta$ and establish k_τ as the number of time intervals from the finer time-scale, which comprise the time period τ. We shall use model (3.18) to analyze further the drift μ_τ and volatility σ_τ of the adjusted returns process $\mathcal{R}_\tau$.

3.4.2 Model of the Relationship

We denote the number of granted patents to the company in one period of time by X_τ and the number of the patent categories spanned by the patent activities of the company during the same period by Y_τ. Additionally, we split the categories into patent categories traditional for the company, denoted by Y_τ^e and new categories, which indicate innovation in the activities of the company. The latter will be denoted by Y_τ^n. Our goal is to study the impact and statistical significance of these factors on the adjusted aggregate return process $\mathcal{R}_\tau$.

We adopt a point of view, in which, we model not only the direct impact of the dynamic factors on the path of the stock return process but we also look at their impact on the features of the process $\{\mathcal{R}_t\}$ as characterized by the drift μ and volatility σ.

We investigate the relation of the processes μ_τ and σ_τ to the processes X_τ, Y_τ^e, and Y_τ^n. To this end, we shall establish a statistical autoregressive model of the following form:

$$\begin{aligned}\mu_\tau = \beta_0^\sigma \sigma_\tau + \beta_0^\mu \mu_{\tau-1} + \\ \beta_1 X_\tau + \beta_2 X_{\tau-1} + \cdots + \beta_\theta X_{\tau-\theta+1} + \\ \beta_1^e Y_\tau^e + \beta_2^e Y_{\tau-1}^e + \cdots + \beta_\theta^e Y_{\tau-\theta+1}^e + \\ \beta_1^n Y_\tau^n + \beta_2^n Y_{\tau-1}^n + \cdots + \beta_\theta^n Y_{\tau-\theta+1}^n + \varepsilon_\tau^\mu.\end{aligned} \tag{3.19}$$

$$\begin{aligned}\sigma_\tau = \alpha_0^\mu \mu_\tau + \alpha_0^\sigma \sigma_{\tau-1} + \\ \alpha_1 X_\tau + \alpha_2 X_{\tau-1} + \cdots + \alpha_\theta X_{\tau-\theta+1} + \\ \alpha_1^e Y_\tau^e + \alpha_2^e Y_{\tau-1}^e + \cdots + \alpha_\theta^e Y_{\tau-\theta+1}^e + \\ \alpha_1^n Y_\tau^n + \alpha_2^n Y_{\tau-1}^n + \cdots + \alpha_\theta^n Y_{\tau-\theta+1}^n + \varepsilon_\tau^\sigma.\end{aligned} \tag{3.20}$$

The equations describe the evolution of the drift and volatility process in response to the vector of weakly exogenous or lagged dependent variables X, Y^e, Y^n. In addition, we analyze the significance of the relation and the number of appropriate delay

terms.

If our analysis encompass a long period of time, the models (3.20) may not be sufficiently accurate. One common approach to this problem is the moving window technique, which is widely used to estimate dynamic changes in factor exposures (see, e.g., [96]). In this case, the model (3.20) will be limited to observations in a portion of the time horizon (a window). In that case, the regression coefficients will change as the portion of data moves from the beginning to the end of the time horizon. The moving window technique has limitations and deficiencies pertaining mainly to the fact that reliable estimates of model parameters can be obtained only if the window is sufficiently large. This approach limits the ability to detect changes that were occurred within a short period of time or very quick, abrupt changes that can occur due to trading. However, our presumption is that the impact of patent activity does not lead to abrupt moves of the stock price but has rather, a longer-term effect. We shall see empirically that this assumption is in-line with the empirical evidence.

A version of the moving window technique is described in [106]. According to this methodology, the regression window that is used for estimation of the model parameters is formed in each point of estimate based on the k-nearest neighbor rule. This method is motivated by the fact that, if no prior information is available, then at any point in the past, observations on both sides of that point are equally important for the estimation process. Therefore, the regression window always includes k observations that are closest in time to the point of estimate. Each data point within the window is assigned a weight decreasing exponentially from the estimation point to both edges of the window. The window is centered around the estimation point, except at the beginning and at the end, where all k-nearest returns are the returns that immediately follow or precede the point of estimate respectively.

Using exponential weights in the moving window technique was suggested in [12] and expanded in [41]. In this approach, the more recent observations have a larger

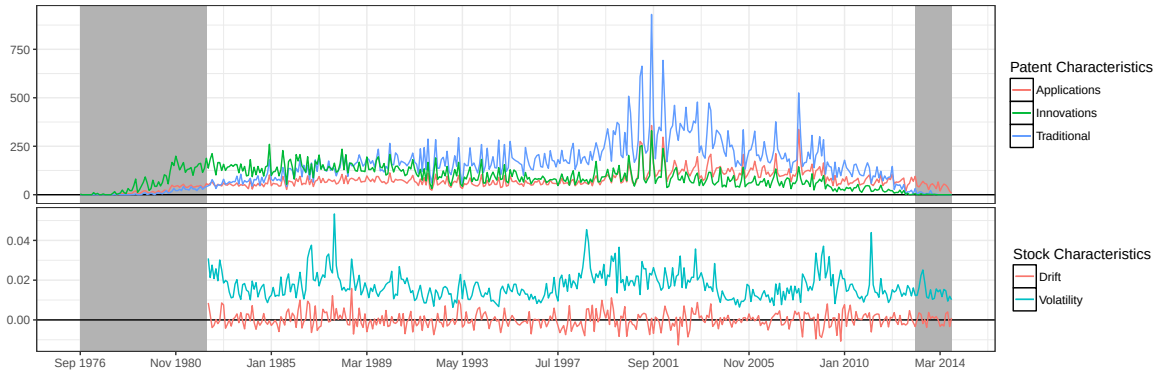

Figure 3.1: Hitachi Ltd. - Patent Activity, Drift, and Volatility

effect than the earlier ones; the weights would gradually decrease within the estimation window from the point of estimate to the end of the window. Furthermore, the weight decrease exponentially with time as each weight is set to be equal to the preceding one multiplied by a number δ, $\delta \in (0, 1)$, which is called a decay factor.

Alternative approach is suggested in [68], where the authors determine dynamic regression coefficients considering the entire data set without the use of predetermined window size for the validity of local models.

Note that we adopt the technique of considering the point of estimation to always be the most recent point of the moving window.

3.4.3 Estimation of Parameters

In order to apply the models described in the previous section, we must compute and estimate the various features characterizing the market-adjusted return process.

To this end, we must first obtain the market-adjusted return process itself. We employ (3.15), where S_t are the focus company's stock prices and I_t are the prices of the Standart & Poor 500 index. We perform the estimation of the market influence coefficient Beta (β) on the entire time horizon. Using this parameter, we obtain the market-adjusted return process R_t.

Using R_t, we perform the estimation of the drift μ_τ and the volatility σ_τ of the aggregate market-adjusted return process $\mathcal{R}_\tau$ in the following way.

Focusing on a time interval $[\kappa * \tau, \kappa * (\tau + 1) - 1]$, we denote $v_i = R_{\kappa * \tau + i}$. We calculate an unbiased estimator for the logarithm of the adjusted returns process:

$$\bar{v}_\tau = \frac{1}{\kappa} \sum_{i=1}^{\kappa} v_i,\ \tau = 1, \ldots, T.$$

The unbiased estimator of the standard deviation for the same quantities is:

$$\sigma_\tau = \sqrt{\frac{1}{\kappa - 1} \sum_{i=1}^{\kappa} \left(v_i - \bar{v}_\tau\right)^2},\ \tau = 1, \ldots, T. \tag{3.21}$$

Due to the adopted model of form (3.12) and the form of the solution (3.13), we can estimate the drift by setting

$$\mu_\tau = \bar{v}_\tau + \frac{1}{2}\sigma_\tau. \tag{3.22}$$

The coefficients in models (3.19) and (3.20) are estimated by using multivariate linear regression on the available data. In a second step, we optimize the model by removing the statistically insignificant regressors. This is accomplished by considering the p-Values of the estimators in conjunction with the Akaiki Information Criterion.

In addition, we employ the moving window technique in order to create a dynamic version of models (3.17), (3.19) and (3.20). The estimation only considers data within a window of size w of the adjusted return process $\mathcal{R}_\tau$. An additional parameter ϖ is introduced controlling the step size of the moving windown. Estimation of the coefficients for the models (3.17), (3.19), and (3.20) is performed while only considering the data available in the window w_i. After estimating the coefficients in first time window w_1, the window is moved by ϖ and another set coefficients is estimated. All models are subjected to the optimization step mentioned previously and are subsequently stored in their entirety, including all data points and residuals.

3.5 Experimental Results

In this section, we present extensive experiments on real-world patent data as well as stock trading data. Our experiments involve several stages of data retrieval and pre-processing before performing model parameter estimations and the statistical analysis of their results.

The first stage of data retrieval consists of the automatic download of stock data and subsequent filtering. In the next phase, all available patent assignment data is parsed for companies of interest. We consolidate the results from the initial parse, generate some meta data in the process and perform additional filtering. We estimate the degree to which the market affects each company and remove this influence from the return process. Considering a set of models for the market-adjusted return process and its characteristics we perform parameter estimation. Finally, we generate a set of models using the moving window technique for a fixed number of lagged patent activity terms. Also, for each set of models, we perform t-test for each of the estimated coefficients.

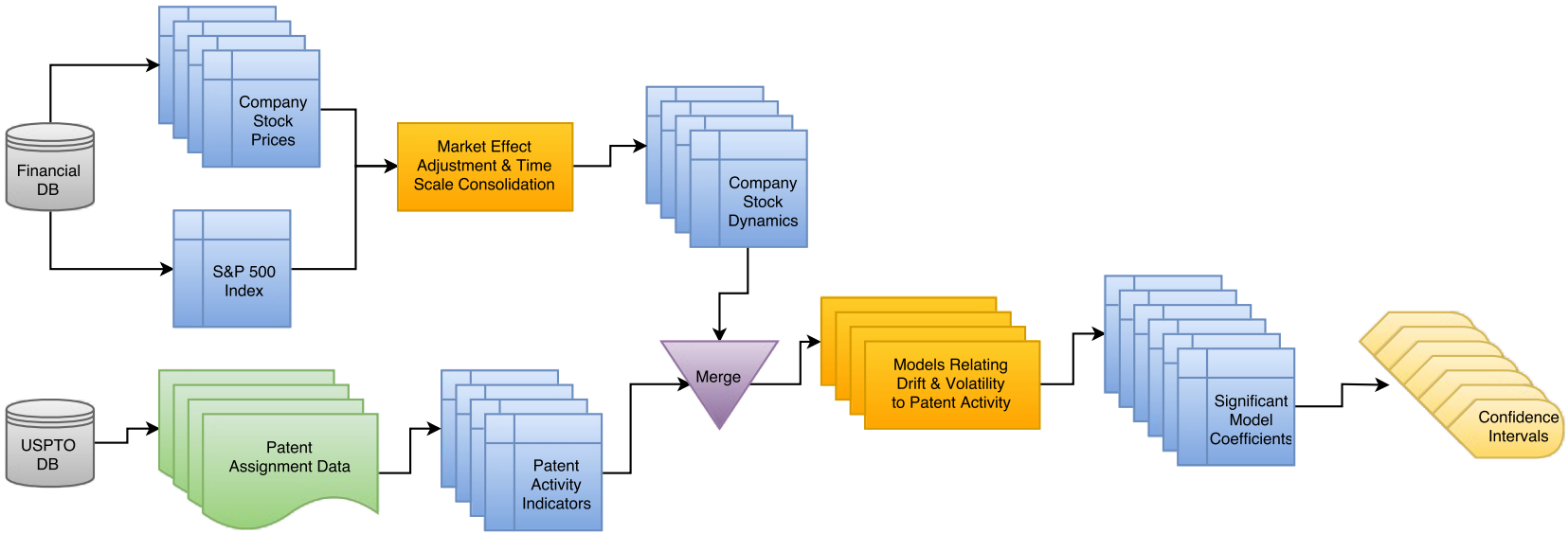

Figure 3.2: Data Flow Diagram – Displays how data is extracted, processed, and combined for modeling and output

3.5.1 Data & Preprocessing

Our focus is on companies in the technology sector, and therefore we use the list of 682 companies included in the NASDAQ as a basis. Additionally, we include two small sets of companies for further support of our claims. The first set of companies (Hitachi Ltd., LG Electronics, Panasonic Corporation, Samsung Electronics Co.) are not traded on the New York Stock Exchange, but are large foreign companies which we consider big players in the technology sector. The second set of companies are being traded on the NYSE, however are not included in the NASDAQ index, these companies (e.g. Amazon.com, Inc., AT&T Inc., eBay Inc., and Sony Corp) are of general interest to us due to their size success. As a result, our initial list of companies comes to a total of 720.

Our initial filtering occurs here, where we ensure that we have sufficient data to perform our analysis. Our estimates indicate that we ought to have approximately ten years of stock trading data in order to generate a sufficiently accurate results. This filtering removes 302 companies and we perform a download of all available daily stock price data for each of the remaining 418 companies.

We retrieved all patent assignment data, which is published by the United States Patent and Trademark Office, and made available as bulk downloads by Google [44]. This data is in the form of compressed XML files containing patent assignment records and totals approximately 30GB when decompressed. Each record may contain multiple assignee fields, all of which are checked in the course of our initial parse. If any of the assignee fields match a company on our list, the parser additionally checks and extracts the unique application, patent, and publication numbers, when available, as well as each of the respective dates. Each patent is associated with one or more classification categories (e.g. "`USER INTERFACE, GUI`") denoted by United States Patent Classification (USPC) codes. These codes are contained in a separate text file which may contain multiple entries for each patent. Every entry contains the patent number followed by the assigned category, thus if a patent is classified

into five categories there will be five entries in the classification file. This single classification file contains more than 32,000,000 entries. The parser creates a separate file for each company and enters each patent assignment record discovered including the collected categories. We extract a total of 775,395 patent assignment records for our 418 companies. Operating under the assumption that a company announces it's discoveries to the public in conjunction with filing for a patent, the date associated with the patent assignment event is the application date. This stage of processing is indeed the most time consuming portion of the entire process.

In the second phase of preprocessing, the parser reads the files generated during the first phase, sorting and aggregating the information; that is, the patent entries for each company are sorted in chronological order since the original entries in the patent assignment records do not appear this way. Intuitively, patent activity does not occur with a fixed time interval between occurrences. Thus the initial data set is event driven rather than a time series, with a consistent scale. In order to remedy this, we create a multivariate time series with a consistent time scale by considering the data on a monthly basis and aggregating the patent events which occurred within a given month. In other words, we count how many patents are filed by the company every month and how many patent categories are covered by these patents; of these categories, how many are new categories, which we will call "innovations," and how many have been previously observed or "traditional." It should be noted, this type of information naturally tapers down in availability the closer we move to the present point in time. This is due to the length of the patent granting process, which averages about two years. Subsequently, we chose to limit the data we use in our experiments to points before January 2013, in order to ensure a high degree of data availability and completeness, as well as avoid any distortions of the statistical models.

Another issue of significance is that some of the companies from the NASDAQ index hold very few, if any, patents. More specifically, we found that of the 418 companies for which we extracted patent-activity data, only 170 companies actually

hold a significant number of patents. Combined, these 170 companies account for 637,478 patents spanning 123,215 unique categories across a total number of 1,589,073 categorical entries.

Although all analysis is conducted for every one of the 170 companies, we cannot possibly display all the results within the page limit allotted. Therefore, we display a representative set of results in Table 3.1 and subsequent figures. The full set of results is made available online.[1]

The series representing patent activity can be seen in the top section of Figure 3.1. In this figure, the two sections of discarded data are shaded grey.

3.5.2 Market Adjustment

To aid us in our investigation, we use the statistical computing platform **R** from this point forward.

In our experiments, we use the adjusted daily closing price of the company stocks from the New York Stock Exchange to calculate daily log returns for the companies under investigation. We have chosen to represent our market the S&P 500 index and we perform the same calculation for its closing price.

In order to maintain a consistent time unit between to the two time series data sets, we choose a monthly time scale, denoted by τ. This choice implies that we have chosen a non-uniform segmentation of the entire time horizon corresponding to the different number of trading days in the different months.

We should note, that the baseline model requires additional special considerations since it still utilizes the daily price data. In order to bring it to a comparable scale to the other two models and facilitate the comparison of the results, we do not compute the daily log returns but rather the monthly log returns. In this sense, the aggregation and unification of time scale for the baseline model happens before removing the market effect. This order of operations provides us with a more accurate aggregate

[1] http://www.constantinevitt.com/patentProject/

process $\mathcal{R}_\tau$ only in the case of the baseline model (3.17). All other estimations, including the estimation of the Beta (β) coefficients is performed on the daily log return process.

Using Equation (3.15), we fit a model for each company, obtaining the Beta (β) coefficients as well as its adjusted return process $\mathcal{R}_\tau$. We proceed by employing standard Q-Q plots, as shown in Figure 3.3, to verify that the observed quantiles are not even close the theoretical quantiles.Indeed, this indicates that there exist external influences on the market-adjusted return process of the stock which we hope to relate to the patent activity.

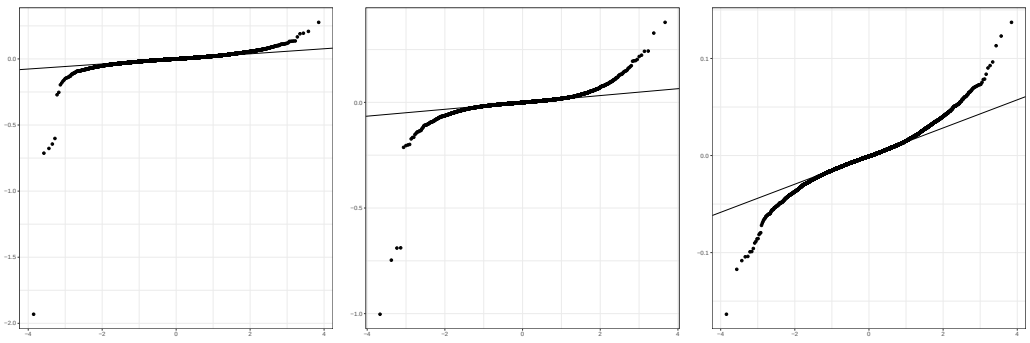

Figure 3.3: Q-Q plots of R_t for Apple, eBay, and Hitachi

Table 3.1 reveals an extremely high correlation of company stock returns compared to the market. Many researchers in quantitative finance recognize the dynamic nature of beta. In the literature, several methods for adjusting betas are avalable, including the most popular ones by [102] and [6].

Name	Symbol	Patents	USPC codes		Market Effect		Models
			Unique	Total	Beta	p-Value	
Adobe Systems Incorporated	ADBE	3033	2837	6943	1.346380675	1.80E-306	133
Advanced Micro Devices, Inc.	AMD	10478	11602	40775	1.558245636	0	154
Apple Inc.	AAPL	9632	7959	16143	1.220812219	1.02E-255	162
Google Inc.	GOOGL	9216	6081	17517	0.920376644	7.62E-152	25
Intel Corporation	INTC	29855	21156	77791	1.326085408	0	172
International Business Machines	IBM	94947	43154	235104	0.982076122	0	233
Microsoft Corporation	MSFT	36279	12691	69038	1.135971032	1.72E-320	136
Nokia Corporation	NOK	14422	8612	25675	1.37439089	4.24E-152	88
Red Hat, Inc.	RHT	1513	1397	2765	1.202637025	3.21E-134	55
Amazon.com, Inc.	AMZN	147	298	526	1.384886922	8.68E-157	69
AT&T Inc.	T	1255	2081	3192	0.746224584	4.09E-262	146
eBay Inc.	EBAY	2185	1169	2706	1.398315771	2.51E-164	61
Sony Corp	SNE	47967	36309	137718	0.906769155	0	206
LG Electronics	066570.KS	24232	16797	47049	0.221982485	4.69E-11	39
Hitachi Ltd.	HTHIY	32980	39916	108521	0.832883463	0	159
Panasonic Corporation	PCRFY	15219	12742	22689	0.724566669	4.78E-119	151
Samsung Electronics Co	005930.KS	82007	43076	171651	0.213280285	2.59E-12	53

Table 3.1: Companies along with their β values

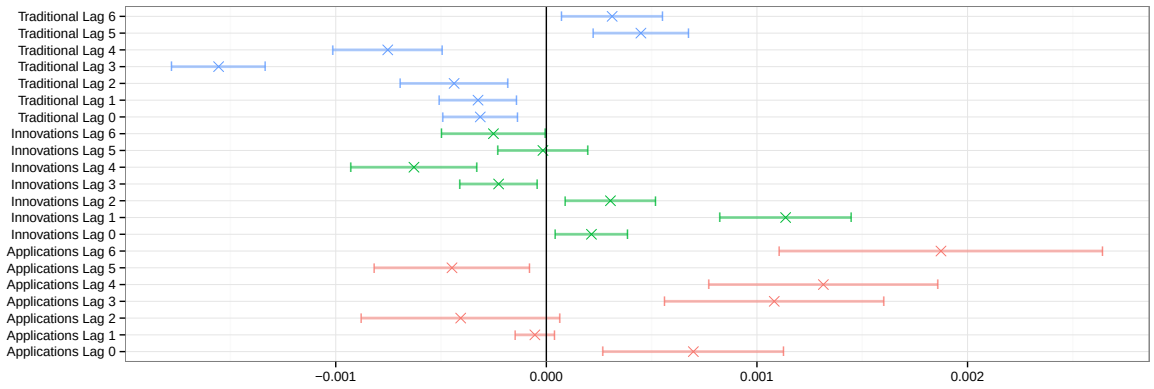

Figure 3.4: Google Inc. - Coefficient sensitivity analysis for (3.17) using 6 lagged terms

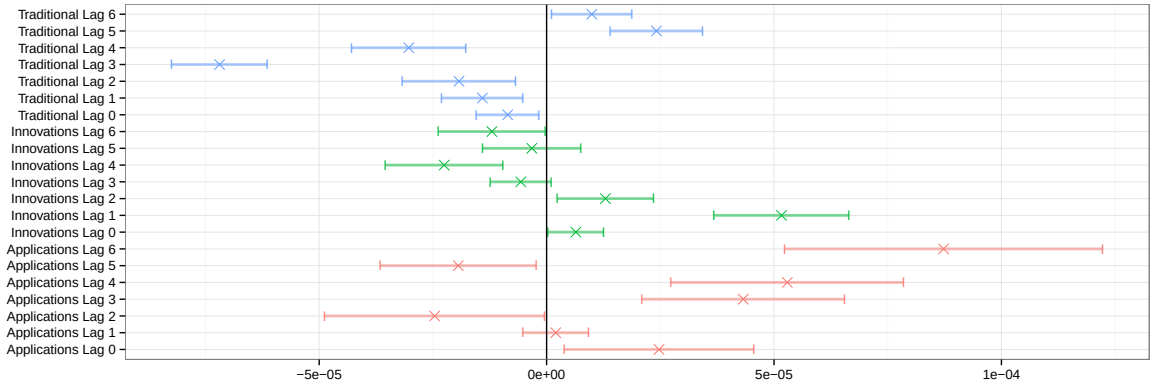

Figure 3.5: Google Inc. - Coefficient sensitivity analysis for (3.19) using 6 lagged terms

The work by [104] contains an extensive comparison of methods of estimation of time-varying betas and filtering techniques, which are commonly used for estimation of individual stock betas. However, our model with a constant beta over the entire time horizon demonstrates extremely high correlation, and thus eliminates the need to create a dynamic model of β for our purposes. Therefore, we have not estimated Beta (β) in a dynamic fashion.

An interesting byproduct of our analysis is the observation of the global impact of the American stock market. We have observed that large companies which are not traded on the New York Exchange, exemplified in this case by Samsung Electronics

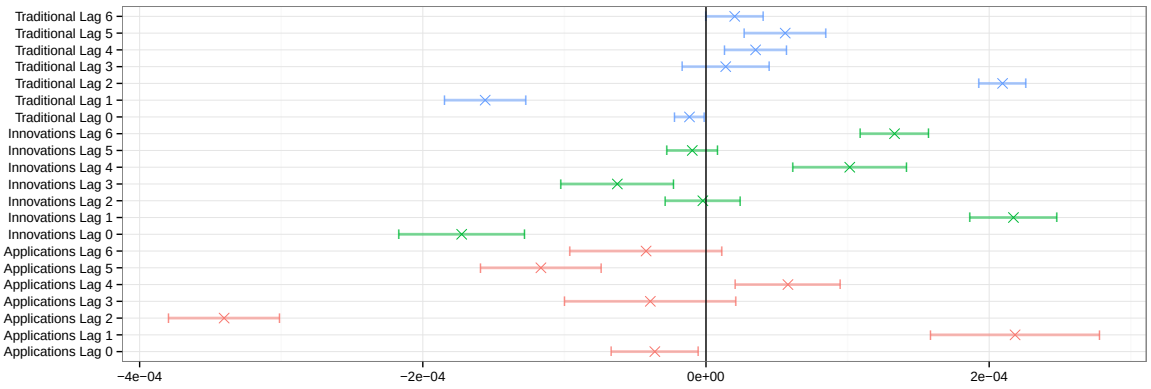

Figure 3.6: Google Inc. - Coefficient sensitivity analysis for (3.20) using 6 lagged terms

and LG Electronics, are substantially influenced by its movement: their beta is smaller than one but still a substantial positive number and their correlation to the Standard & Poor Index is significant.

We apply Equations (3.22) and (3.21) to the market-adjusted return process R_t to obtain the monthly drift and volatility estimates μ_τ and σ_τ while considering τ to be one month. The resulting series, illustrated by data for Hitachi Ltd., can be seen in the lower part of Figure 3.1.

3.5.3 Model Fitting

Our model fitting procedure has three primary parameters in addition to the input data. In the experiments, we consider a window size w of 48 months and a step size ϖ of 3 months. Additionally, we consider an array of options for the number of lagged terms to include in the model fitting. We apply the models in an iterative manner, using a fixed number of lagged patent activity terms, moving the window by ϖ until we traverse the entire set of available data for the focus company. On the first set of iterations we consider the patent activity from current month as well as the most recent 6 months, thus 7 observations on each of the model parameters X_τ number of patents, Y_τ^e number of traditional categories, and Y_τ^n the number of

innovative categorie spanned by the patent activities. The resulting models (3.17) and (3.19) have a total of 22 factors when considering the one autoregressive term in them. The volatility model (3.20) has an additionally term for the current drift, thus totaling 23 factors. After the initial parameter estimation during each iteration, the models go through a recursive subroutine being reduced and further optimized by excluding statistically insignificant regressors. As previously mentioned, we leverage the p-Values of the estimated coefficients in conjunction with the Akaiki Information Criterion for the model in order to determine the best combination factors. This process leads to a relatively wide variety of models with varying numbers of significant regression terms. Therefore, once the available data is traversed for the chosen number of lagged terms, we collect the estimated coefficients. The coefficients from a given model are recorded as a vertical vector in an m by n matrix C where m is total number of coefficients in the set of models before reduction and n are the number of models.

In the case where a coefficient i from a model j has been eliminated during the model reduction phase, we insert zero for c_{ij}. After completing this procedure for every lag term option in the array and model formulation, we move on to the next company and restart the model fitting procedure. The matrix C is collected and written to a file for every model type, every company, and every fixed lag term option. We have considered between six and twelve lagged patent activity terms. In our case, since we have considered between six and twelve lagged patent activity terms and three model types, we generate 21 output files per company.

We are able to establish statistically significant models of all three forms (3.17), (3.19), and (3.20) for all of the companies under consideration. However, the estimated coefficients are company specific and do not seem consistent accross companies in the same time window. Further analysis may try to establish whether there is leader-follower effect responsible for this inconsistency. The impact of patent activities as reflected in the model coefficients also changes in time when following a

specific company over the time horizon in our study. More detailed comments in that regard are presented in the next section.

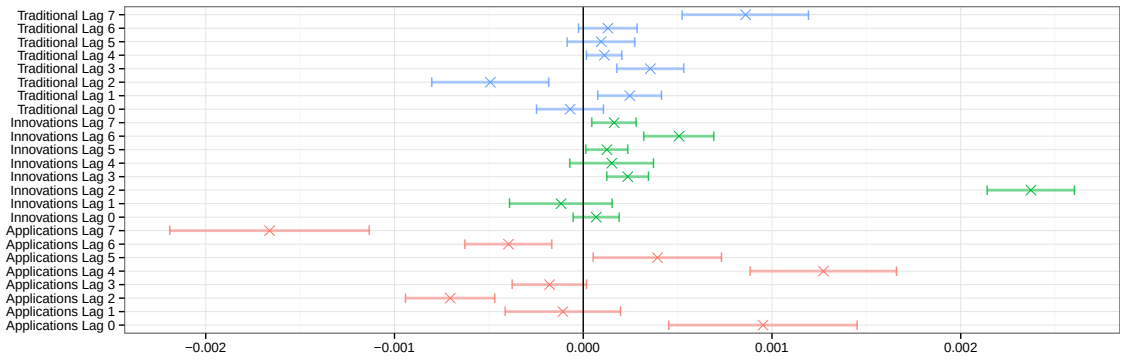

Figure 3.7: Model coefficients for eBay with 7 sets of lagged terms

3.5.4 Impact Significance

An illustration of the estimates and their sensitivity analysis can be seen displayed in Figure 3.7, for our eBay example using 7 lagged terms and showcasing the volatility model formulation (3.20). Figures 3.4, a3.5, and 3.6 show the same analysis applied to all three models types (3.17), (3.19) and (3.20) when applied to Google's data.

Recall that the columns of matrix C contain the estimated model coefficients for each time window. We would like to analyze the statistical significance on the sample of estimated coefficients for each model type, number of lagged terms, and company. We perform a t-test on the rows of matrix C, thus obtaining a mean estimate and the 95% confidence interval for each model coefficient. The significance of a factor over the entire time horizon is evidenced by the exclusion of zero from its confidence interval.

Taking a closer look at Figure 3.7, we observe that eBay's innovations from two months prior have significantly positive impact on the current volatility, while traditional category contributions as well as applications from two months prior have a significantly negative impact on the current volatility. In other words, the results

indicate that the applications and traditional categories help stabilize the price, while innovations spark investor interest and induce speculation. It should be noted that the autoregressive volatility term from model (3.20) is always present with a positive relatively large coefficient and thus for the sake of preserving the scale of the graph is not included. Similarly, the current drift term from model (3.20) is also excluded from the graphic for the sake of scale, despite being highly significant with a negative coefficient.

Figures 3.4, 3.5, and 3.6 showcase a recurring theme in our results. We can clearly observe that the coefficients for the formulations (3.17) and (3.19) have a very similar pattern. Their analysis is displayed in the top and middle graphs respectively. This is natural and expected, due to the nature of the two processes being modeled. However, the coefficients for (3.20) display a very different pattern. The latter pattern is impossible to infer from the standard approach to factor investigation, relating the process itself to the factors. Our observations only came to light due to the innovative approach of relating the stock volatility directly to the patent-activity indicators.

The impact of the factors varies across companies. This may be an indicator that a non-linear relation between the volatility and the patent activity exists.

3.6 Concluding Remarks

In this thesis, we provided a focused study of the relationship between the patent activities of high-tech companies and the dynamics of their stock price movement. To the best of our knowledge, we are the first to propose a model for relating patent data mining and financial data modeling. Along this line, our findings indicate that the patent activities of the company play a role as a latent factor of the stock process. By exploiting a stock-price model based on stochastic differential equations, we have revealed that the chosen patent-activity indicators have significant impact on the process itself, as well as on its drift and volatility. The introduction of the new

approach of modeling the drift and volatility directly in addition to the adjusted returns, allows us to establish the influence of these factors on the most important features of the adjusted stock returns, which would be otherwise hard to extract. The proposed approach of modeling the impact of the factors directly on the features of the process may be of independent interest when relating dynamic data streets and looking to establish relations. Furthermore, our results have implied that the factors pertaining to number of patent categories are most influential. This may indicate that the company has created and patented influential technology relevant for many business activities. The patent activity of this type has a bigger impact in the business area of this company, which results in a bigger movement in the market, thus, manifesting in a bigger volatility. Moreover, the presence of different signs of the regressor coefficients may indicate that the effect of patent activity is different depending on the relation of the drift to its expected value. This may also indicate that a non-linear models including cross terms may be relevant.

Chapter 4
Future work

In this thesis, we have explored recent developments of risk theory and have applied two recently introduced models of risk: coherent measures of risk and a basic stochastic differential equation for stock returns. We have applied these models to two areas of machine learning: classification and identification of the impact of patent activity on the stock-price dynamics of high-tech companies. Our theoretical analysis and numerical experiments demonstrate that new insight is provided when this type of models are used. In the course of our investigation, we have encountered a variety of questions, some of which go beyond this thesis and are subject of future research. We mention here some of the most closely related questions that we find of interest. Of course, other risk models may prove relevant to these or other problems of machine learning.

4.1 Risk-averse Classification Methods

In the context of classification, it is desirable to explore the effect of using higher order measures of risk. In the reported experiments, we have used only first order measures of risk: the average value at risk and the mean-semideviation of order one in combination with the expected value. Using higher order (dual) measures of risk or mean-semi-deviation of higher order requires the application of numerical methods for general non-linear optimization.

Furthermore, one has to take into account the effect of a more general non-linear objective on the numerical method for solving the RSSVM with normalization of

the classifier. The method needs more extensive exposure to different data sets and comparison with a larger variety of classification techniques.

The implementation of the risk-averse approach on multiple classes is another challenge given the very large set of possible risk measures. When dealing with very large datasets with many classes, we may consider non-linear scalarizations rather than linear one for obtaining an efficient risk allocation. This approach may prove more efficient numerically. Larger data sets would require a distributed implementation of the proposed classification methodology.

Further theoretical analysis of the asymptotic behavior of the risk-averse classifier is of interest, which would allow us to obtain a confidence region of the classifier. To that effect, new research in the area optimization problems with coherent measures of risk is necessary, which would address statistical inference of problems involving multiple populations.

4.2 The Impact of Patent Activity

In the context of patent activity and its impact on the dynamics of the stock returns, one may explore other stochastic models for the market-adjusted return process $\{R_t, t \geq 0\}$. One possibility would be to adopt regime-switching models for the marked-adjusted return, as well as for its drift and volatility. This approach may be very relevant particularly to the influence of patent activity on the volatility, since that impact may be different in the context of positive and negative drift. Statistical test for change point detection and regime switching would facilitate discovery of the point of impact of specific patent activity.

Given the importance of patent activity in the technology sector and our current findings, an autoregressive model of the drift μ_τ and a threshold model for the variance may provide new insight into the impact of patent activity on the company's stock price.

Another unexplored avenue is the adoption of heavy tailed process models such as Paretian-type models as described in [83, 82]. Many experts in mathematical finance claim that models with heavier tails than the normal distribution are more relevant and fit better to the dynamics of the stock prices or stock returns, respectively. In our context, we could fit a heavy-tailed model for the process R_τ and test statistically whether that model represents the process in a better way. By adopting this point of view, it is possible to have a new insight into the impact of patent activity on the stock movement.

References

[1] Carlo Acerbi and Dirk Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002.

[2] Barry C Arnold. The lorenz order in the space of distribution functions. In *Majorization and the Lorenz Order: A Brief Introduction*, pages 29–44. Springer, 1987.

[3] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[4] Gerald Beer. *Topologies on closed and closed convex sets*, volume 268. Springer Science & Business Media, 1993.

[5] Irad Ben-Gal. Outlier detection. In *Data mining and knowledge discovery handbook*, pages 131–146. Springer, 2005.

[6] Marshall E Blume. Betas and their regression tendencies. *The Journal of Finance*, 30(3):785–795, 1975.

[7] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986.

[8] Tim Bollerslev, Robert F Engle, and Jeffrey M Wooldridge. A capital asset pricing model with time-varying covariances. *The Journal of Political Economy*, pages 116–131, 1988.

[9] J Frédéric Bonnans and Alexander Shapiro. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

[10] Dragos Bozdog, Ionut Florescu, Khaldoun Khashanah, and Jim Wang. Rare events analysis for high-frequency equity data. *Wilmott*, 2011(54):74–81, 2011.

[11] Michael PS Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terrence S Furey, Manuel Ares, and David Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–267, 2000.

[12] Robert G Brown. Exponential smoothing for predicting demand. In *Operations Research*, volume 5, pages 145–145. Inst. Operations Research Management Sciences, 1957.

[13] Anna L Buczak and Erhan Guven. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2):1153–1176, 2016.

[14] Howard S Burkom, Sean Patrick Murphy, and Galit Shmueli. Automated time series forecasting for biosurveillance. *Statistics in medicine*, 26(22):4202–4218, 2007.

[15] Ismail Butun, Salvatore D Morgera, and Ravi Sankar. A survey of intrusion detection systems in wireless sensor networks. *IEEE communications surveys & tutorials*, 16(1):266–282, 2014.

[16] Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets, 1984.

[17] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[18] Jie Chen, Hongxing He, Graham Williams, and Huidong Jin. Temporal sequence associations for rare events. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 235–239. Springer, 2004.

[19] Nai-Fu Chen, Richard Roll, and Stephen A Ross. Economic forces and the stock market. *Journal of business*, pages 383–403, 1986.

[20] Patrick Cheridito and Tianhui Li. Risk measures on orlicz hearts. *Mathematical Finance*, 19(2):189–214, 2009.

[21] Bill Chiu, Eamonn Keogh, and Stefano Lonardi. Probabilistic discovery of time series motifs. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–498. ACM, 2003.

[22] D. R. Cox and Peter A. W. Lewis. *The statistical analysis of series of events [by] D.R. Cox and P.A.W. Lewis*. Methuen London, 1966.

[23] Brian Culshaw and José Miguel López-Higuera. Engineering a high-tech business: Entrepreneurial experiences and insights. SPIE, 2008.

[24] Justin R Davis and Stan Uryasev. Analysis of tropical storm damage using buffered probability of exceedance. *Natural Hazards*, 83(1):465–483, 2016.

[25] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[26] Darinka Dentcheva and Gabriela Martinez. Two-stage stochastic optimization problems with stochastic ordering constraints on the recourse. *European Journal of Operational Research*, 219(1):1–8, 2012.

[27] Darinka Dentcheva and Spiridon Penev. Shape-restricted inference for lorenz curves using duality theory. *Statistics & probability letters*, 80(5):403–412, 2010.

[28] Darinka Dentcheva, Spiridon Penev, and Andrzej Ruszczyński. Kusuoka representation of higher order dual risk measures. *Annals of Operations Research*, 181(1):325–335, 2010.

[29] Darinka Dentcheva, Spiridon Penev, and Andrzej Ruszczyński. Statistical estimation of composite risk functionals and risk optimization problems. *Annals of the Institute of Statistical Mathematics*, pages 1–24, 2016.

[30] A Rogier T Donders, Geert JMG Van Der Heijden, Theo Stijnen, and Karel GM Moons. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091, 2006.

[31] Laurent El Ghaoui, Gert René Georges Lanckriet, Georges Natsoulis, et al. *Robust classification with interval data*. Computer Science Division, University of California, 2003.

[32] Graham Elliott and Allan Timmermann. *Handbook of economic forecasting*. Elsevier, 2013.

[33] Robert F Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007, 1982.

[34] Ugo Fano. Ionization yield of radiations. ii. the fluctuations of the number of ions. *Physical Review*, 72(1):26, 1947.

[35] I Florescu, F Viens, and Maria C Mariani. Handbook of modeling high-frequency data in finance. 2012.

[36] Hans Föllmer and Alexander Schied. *Stochastic finance: an introduction in discrete time*. Walter de Gruyter, 2011.

[37] Joseph L Gastwirth. The estimation of the lorenz curve and gini index. *The Review of Economics and Statistics*, pages 306–316, 1972.

[38] Pierre Geurts. Pattern extraction for time series classification. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 115–127. Springer, 2001.

[39] Mohammad Al Hasan, W. Scott Spangler, Thomas Griffin, and Alfredo Alba. Coa: Finding novel patents through text analysis. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 1175–1184, New York, NY, USA, 2009. ACM.

[40] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York, 2009.

[41] CC Holt. Forecasting trends and seasonal by exponentially weighted moving averages. *ONR Memorandum*, 52, 1957.

[42] James Honaker and Gary King. What to do about missing values in time-series cross-section data. *American Journal of Political Science*, 54(2):561–581, 2010.

[43] Peter J Huber. *Robust statistics*. Springer, 2011.

[44] Google Inc. United states patent and trademark office bulk downloads, 2015. [Online; accessed 30-January-2015].

[45] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[46] Dimitrije Jankov, Sourav Sikdar, Rohan Mukherjee, Kia Teymourian, and Chris Jermaine. Real-time high performance anomaly detection over data streams: Grand challenge. In *Proceedings of the 11th ACM International Conference on Distributed and Event-based Systems*, pages 292–297. ACM, 2017.

[47] Xin Jin, S. Spangler, Ying Chen, Keke Cai, Rui Ma, Li Zhang, X. Wu, and Jiawei Han. Patent maintenance recommendation with patent information network model. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 280–289, Dec 2011.

[48] Elyes Jouini, Walter Schachermayer, and Nizar Touzi. Optimal risk sharing for law invariant monetary utility functions. *Mathematical Finance*, 18(2):269–292, 2008.

[49] Eamonn Keogh and Chotirat Ann Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and information systems*, 7(3):358–386, 2005.

[50] Eamonn J Keogh and Michael J Pazzani. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In *Kdd*, volume 98, pages 239–243, 1998.

[51] Masaaki Kijima and Masamitsu Ohnishi. Mean-risk analysis of risk aversion and wealth effects on optimal portfolios with multiple investment opportunities. *Annals of Operations Research*, 45(1):147–163, 1993.

[52] Dohyun Kim, Bangrae Lee, Hyuck Jai Lee, Sang Pil Lee, Yeongho Moon, and M.K. Jeong. Automated detection of influential patents using singular values. *Automation Science and Engineering, IEEE Transactions on*, 9(4):723–733, Oct 2012.

[53] P.E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2010.

[54] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[55] Pavlo A Krokhmal. Higher moment coherent risk measures. 2007.

[56] Shigeo Kusuoka. On law invariant coherent risk measures. In *Advances in mathematical economics*, pages 83–95. Springer, 2001.

[57] Gert RG Lanckriet, Laurent El Ghaoui, Chiranjib Bhattacharyya, and Michael I Jordan. A robust minimax approach to classification. *The Journal of Machine Learning Research*, 3:555–582, 2003.

[58] Michael Landsberger and Isaac Meilijson. Co-monotone allocations, bickel-lehmann dispersion and the arrow-pratt measure of risk aversion. *Annals of Operations Research*, 52(2):97–106, 1994.

[59] T Warren Liao. Clustering of time series dataa survey. *Pattern recognition*, 38(11):1857–1874, 2005.

[60] M. Lichman. UCI machine learning repository, 2013.

[61] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 333. John Wiley & Sons, 2014.

[62] Shinhua Liu. Index futures and predictability of the underlying stocks returns: The case of the nikkei 225. *Journal of Financial Services Research*, 34(1):77–91, 2008.

[63] Yan Liu, Pei-yun Hseuh, Rick Lawrence, Steve Meliksetian, Claudia Perlich, and Alejandro Veen. Latent graphical models for quantifying and predicting patent quality. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1145–1153, New York, NY, USA, 2011. ACM.

[64] O. Lorenz. Methods of measuring concentration of wealth. *Journal of the American Statistical Association*, 9:209–219, 1905.

[65] Michael Ludkovski and Ludger Rüschendorf. On comonotonicity of pareto optimal risk sharing. *Statistics & Probability Letters*, 78(10):1181–1188, 2008.

[66] Junshui Ma and Simon Perkins. Online novelty detection on temporal sequences. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–618. ACM, 2003.

[67] Yifei Ma, Li Li, Xiaolin Huang, and Shuning Wang. Robust support vector machine using least median loss penalty. In *the 16th IFAC World Congress*, pages 11208–11213, 2011.

[68] Michael Markov, Vadim Mottl, and Ilya Muchnik. Principles of nonstationary regression estimation: A new approach to dynamic multi-factor models in finance. *DIMACS, Rutgers Univ., Piscataway, NJ, Tech. Rep*, 47, 2004.

[69] Kaisa Miettinen. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media, 1999.

[70] Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 525–531. IEEE, 2001.

[71] M. Musiela and M. Rutkowski. *Martingale Methods in Financial Modelling*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2010.

[72] Yoshifumi Nakata, Michael R Ward, and Xingyuan Zhang. The market value of patenting: New findings from japanese firm level data. 2011.

[73] Victor Ng, Robert F Engle, and Michael Rothschild. A multi-dynamic-factor model for stock returns. *Journal of Econometrics*, 52(1):245–266, 1992.

[74] W. Ogryczak and A. Ruszczyński. Dual stochastic dominance and related mean risk models. *SIAM Journal on Optimization*, 13:60–78, 2002.

[75] Włodzimierz Ogryczak and Andrzej Ruszczyński. From stochastic dominance to mean-risk models: Semideviations as risk measures1. *European Journal of Operational Research*, 116(1):33–50, 1999.

[76] WLodzimierz Ogryczak and Andrzej Ruszczynski. Dual stochastic dominance and related mean-risk models. *SIAM Journal on Optimization*, 13(1):60–78, 2002.

[77] Rulin Ouyang, Liliang Ren, Weiming Cheng, and Chenghu Zhou. Similarity search and pattern discovery in hydrological time series data mining. *Hydrological processes*, 24(9):1198–1210, 2010.

[78] John Paparrizos and Luis Gravano. Fast and accurate time-series clustering. *ACM Transactions on Database Systems (TODS)*, 42(2):8, 2017.

[79] Darshak Patel and Michael R Ward. Using patent citation patterns to infer innovation market competition. *Research Policy*, 40(6):886–894, 2011.

[80] Ivan Popivanov and Renee J Miller. Similarity search over time-series data using wavelets. In *Data Engineering, 2002. Proceedings. 18th International Conference on*, pages 212–221. IEEE, 2002.

[81] Bo Qian and Khaled Rasheed. Foreign exchange market prediction with multiple classifiers. *Journal of Forecasting*, 29(3):271–284, 2010.

[82] Christian Menn Rachev, Svetlozar and Frank J. Fabozzi. *Fat-tailed and skewed asset return distributions; implications for risk management, portfolio selection, and option pricing*. John Willey & Sons, 2005.

[83] Svetlozar Rachev and Stefan Mittnik. *Stable Paretian Models in Finance*. John Willey & Sons, New York, 2000.

[84] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[85] R Tyrrell Rockafellar, Stan Uryasev, and Michael Zabarankin. Generalized deviations in risk analysis. *Finance and Stochastics*, 10(1):51–74, 2006.

[86] R Tyrrell Rockafellar, Stan Uryasev, and Michael Zabarankin. Risk tuning with generalized linear regression. *Mathematics of Operations Research*, 33(3):712–729, 2008.

[87] R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.

[88] R.T. Rockafellar. *Conjugate Duality and Optimization*. Regional Conference Series in Applied Mathematics, SIAM, Philadelphia, 1974.

[89] A Rozhkov. Forecasting russian stock market trends. *Problems of Economic Transition*, 48(6):48–65, 2005.

[90] Andrzej Ruszczynski and Alexander Shapiro. Optimization of convex risk functions. *Mathematics of operations research*, 31(3):433–452, 2006.

[91] D. Sculley, Matthew Eric Otey, Michael Pohl, Bridget Spitznagel, John Hainsworth, and Yunkai Zhou. Detecting adversarial advertisements in the wild. In *Proceedings of the 17th ACM SIGKDD International Conference on Data Mining and Knowledge Discovery*, 2011.

[92] M Shaked and J Shantikumar. Stochastic orderings, 2007.

[93] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

[94] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*, volume 16. SIAM, 2014.

[95] William F Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk*. *The journal of finance*, 19(3):425–442, 1964.

[96] William F Sharpe. Asset allocation: Management style and performance measurement. *The Journal of Portfolio Management*, 18(2):7–19, 1992.

[97] Meng-Jung Shih, Duen-Ren Liu, and Ming-Li Hsu. Mining changes in patent trends for competitive intelligence. In Takashi Washio, Einoshin Suzuki, KaiMing Ting, and Akihiro Inokuchi, editors, *Advances in Knowledge Discovery and Data Mining*, volume 5012 of *Lecture Notes in Computer Science*, pages 999–1005. Springer Berlin Heidelberg, 2008.

[98] S. Spangler, Ying Chen, J. Kreulen, S. Boyer, T. Griffin, A Alba, L. Kato, A Lelescu, and Su Yan. Simple: Interactive analytics on patent data. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 426–433, Dec 2010.

[99] Jie Tang, Bo Wang, Yang Yang, Po Hu, Yanting Zhao, Xinyu Yan, Bo Gao, Minlie Huang, Peng Xu, Weichang Li, and Adam K. Usadi. Patentminer: Topic-driven patent analysis and mining. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1366–1374, New York, NY, USA, 2012. ACM.

[100] Edward R Tufte. The visual display of quantitative information. *Journal for Healthcare Quality*, 7(3):15, 1985.

[101] Alireza Vahdatpour, Navid Amini, and Majid Sarrafzadeh. Toward unsupervised activity discovery using multi-dimensional motif detection in time series. In *IJCAI*, volume 9, pages 1261–1266, 2009.

[102] Oldrich A Vasicek. A note on using cross-sectional information in bayesian estimation of security betas. *The Journal of Finance*, 28(5):1233–1239, 1973.

[103] Mark W Watson and Robert F Engle. Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, 23(3):385–400, 1983.

[104] Curt Wells. *The Kalman filter in finance*, volume 32. Springer, 1996.

[105] Michael A Wiles, Neil A Morgan, and Lopo L Rego. The effect of brand acquisition and disposal on stock returns. *Journal of Marketing*, 76(1):38–58, 2012.

[106] Viktor Zurakhinsky. Capturing changes in style exposure. *The Journal of Performance Measurement*, pages 48–50, 1997.