2017

Essays on risk management in portfolio optimization and gas supply networks

Onur Babat
Lehigh University

Essays on risk management in portfolio optimization and gas supply networks

by

Onur Babat

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Doctor of Philosophy

in

Industrial and Systems Engineering

Lehigh University

(May, 2017)

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Date

Dissertation Advisor:

_____
Dr. Luis F. Zuluaga

_____
Accepted Date

Committee Members:

_____
Dr. Luis F. Zuluaga, Committee Chair

_____
Dr. Robert Storer

_____
Dr. Aurelie Thiele

_____
Dr. Ali Esmaili

_____
Dr. Juan C. Vera

Acknowledgements

First and foremost, I would like to express my gratitude to my advisor, Dr. Luis Zuluaga, whose expertise, understanding, and patience, added considerably to my graduate school experience. I appreciate his vast knowledge and skill, and his great assistance in writing my dissertation.

I would like to thank the other members of my committee, Dr. Robert Storer, Dr. Aurelie Thiele, Dr. Ali Esmaili and Dr. Juan Vera for the assistance they provided at all levels of my dissertation. I would like to thank Dr. Camilo Mancilla, Dr. Pete Verderame and Dr. Erdem Arslan for providing insightful comments and constructive advice during my internship at Air Products and Chemicals, Inc. I offer my sincerest gratitude to Dr. Tamas Terlaky for his continued support during my Ph.D. I am also thankful to Rita Frey and Kathy Rambo, who always help me to go through all administration challenges. I would like to thank Dr. Kathleen Hutnik and Amy McCrae for being such good friends and making my life a lot easier with their support during my doctoral journey. Also, I thank all my friends at Lehigh University for their joyful, and hopefully lifelong, companion.

I would like to dedicate this dissertation to my family for the support they provided me through my entire life. My sister, Melike, for her never ending love and support. And my parents, Seval and Cemal, for all selfless support and sacrifices they made for me. I owe them tons for igniting and encouraging my love for learning and pursuit of knowledge from a young age.

Contents

List of Tables

List of Figures

Abstract

This work focuses on developing algorithms and methodologies to solve problems dealing with uncertainty in portfolio optimization and industrial gas networks.

First, we study the Mean-SemiVariance Project (MSVP) portfolio selection problem, where the objective is to obtain the optimal risk-reward portfolio of non-divisible projects when the risk is measured by the semivariance of the portfolio's Net-Present Value (NPV) and the reward is measured by the portfolio's expected NPV. Similar to the well-known Mean-Variance portfolio selection problem, when integer variables are present (e.g., due to transaction costs, cardinality constraints, or asset illiquidity), the MSVP problem can be solved using Mixed-Integer Quadratic Programming (MIQP) techniques. However, conventional MIQP solvers may be unable to solve large-scale MSVP problem instances in a reasonable amount of time. In this paper, we propose two linear solution schemes to solve the MSVP problem; that is, the proposed schemes avoid the use of MIQP solvers and only require the use of Mixed-Integer Linear Programming (MILP) techniques. In particular, we show that the solution of a class of real-world MSVP problems, in which project returns are positively correlated, can be accurately approximated by solving a single MILP problem. In general, we show that the MSVP problem can be effectively solved by a sequence of MILP problems, which allow us to solve large-scale MSVP problem instances faster than using MIQP solvers. We illustrate our solution schemes by solving a real MSVP problem arising in a Latin American oil and gas company. Also, we solve instances of the MSVP problem that

are constructed using data from the PSPLIB library of project scheduling problems. Both approaches are empirically shown to be effective and outperforming the default benchmark MIQP solver to find near-optimal solutions for the selected instances of the MSVP problem (Sefair et al. 2017).

Second, we present an algorithm to compute near-optimal Value-at-Risk (VaR) portfolios. It is known to be difficult to compute optimal VaR portfolios; that is, an optimal risk-reward portfolio allocation using VaR as the risk measure. This is due to VaR being non-convex and of combinatorial nature. In particular, it is well-known that the VaR portfolio problem can be formulated as a mixed-integer linear program (MILP) that is difficult to solve with current MILP solvers for medium to large-scale instances of the problem. The proposed algorithm addresses the shortcomings of the MILP formulation in terms of solution time. To illustrate the efficiency of the presented algorithm, numerical results are presented using historical asset returns from the US financial market. Empirical results suggest that the developed algorithm obtaining a lower bound for VaR outperforms the recently proposed algorithms from the literature. Additionally, we also show that the developed algorithms are able to obtain and guarantee near-optimal solutions for large scale instances of VaR portfolio optimization problem more efficiently than the off the shelf commercial solvers within 1% accuracy (Babat et al. 2017b).

Third, we analyze the impact of the sensor reading errors on parameters that affect the production costs of a leading US industrial gas supply company. For this purpose, a systematic methodology is applied first to determine the relationship between the system output and input parameters, and second to identify the assigned input sensors whose readings need to be improved in a prioritized manner based on the strength of those input-output relationships. The two main criteria used to prioritize these sensors are the decrease in production costs and the decrease in production costs volatility obtained when the selected sensors precision is improved. To illustrate the

effectiveness of the proposed approach, we first apply it to a simplified version of the real supply network model where the results can be readily validated with the simulated data. Then, we apply and test the proposed approach in the real supply network model with historical data. The experiments suggest that we are able to obtain a significant decrease in production costs and in production costs volatility by prioritizing the sensors' maintenance subject to a limited budget (Babat et al. 2017a).

Finally, we analyze the performance of portfolio allocation strategies using clustering techniques based on financial asset's correlation matrices. The Markowitz's mean-variance framework uses first and second order sample moment estimators which are highly subject to estimation errors. The estimation error on the moments could be very significant and it may offset the benefits obtained from the diversification of the portfolio. There are a number of methodologies proposed in the literature to reduce the effect of the estimation error on the moment estimators. A group of these are based on the clustering approaches using sample correlation coefficients as the similarity measure. The idea is to obtain a hierarchical structure between the financial assets and then to use this information to filter the underlying true representative economic information between the assets and to reflect it in a modified correlation matrix. The objective of this study is to replicate and verify some of the published work comparing different allocation strategies and also incorporating recently published hierarchical clustering based portfolio selection strategies into out of sample performance evaluation across different datasets. Initial findings suggest that the difference between the performance of the classical strategies and the recently developed clustering based methodologies are not statistically significant from each other when only positive weights are allowed in the portfolios.

Chapter 1

Introduction

Uncertainty is a situation that implies the imperfect knowledge and information. On the other hand, risk is a state of uncertainty where some possible outcomes have an undesired effect or loss. All real world projects involve both uncertainty and risk, and uncertainty and risk involve both threat and opportunity. The recognition of them could yield a more desirable and appropriate level of benefit in return for the resource commitment. However, uncertainty quantification could also be very challenging to address in both computational and real world applications.

Uncertainty arises in different ways in almost any field, including insurance, statistics, economics, finance, engineering, and information science (cf. (Cairns 2000, Shackle 1955, Li 2004, Kline 1985)). Making decisions under uncertainty carries an intrinsic risk. Modeling and controling risk and uncertainty is a very challenging problem. This is even more challenging when combinatorial constraints are present (e.g., lot sizing constraints, binary decisions, etc.). The first two chapters in this study specifically focus on obtaining effective solution schemes based on Integer Programing (IP) techniques to solve portfolio optimization problems under stochastic and integer constraints.

The selection of the best investment projects within a set of alternatives is crucial

to any firm facing competition. Moreover, the ability to build portfolios that efficiently allocate scarce resources contributes to the achievement of corporate goals in the long run. Typically, a portfolio's expected profit is considered the single most important corporate goal to be maximized; however, it is not the only one: the fitness of a firm's portfolio should also involve a measure of the portfolio's volatility or risk. For instance, a portfolio with very attractive expected profits might expose the company to a large loss with high probability, whereas a low-risk portfolio might secure the company lower but more certain profits. For these reasons, the problem of selecting assets to create an optimal risk-reward portfolio has been actively considered in the literature. (cf. (Markowitz 1952, Rockafellar and Uryasev 2000, Konno and Yamazaki 1991, Wang and Hwang 2007)).

In his seminal work, (Markowitz 1952), proposed a risk-reward framework a single period model of investment; since then variance has been widely accepted as a risk measure, except that it is questionable to use variance as a measure of risk. Although, for both theoretical and computational reasons, variance has been extensively used in the literature and in practice, academics and practitioners have developed down-side risk measures such as semivariance, Value-at-Risk (VaR), Conditional-Value-at-Risk (CVaR) that penalize extreme losses. These measures of risk take into account the skewness of real asset returns, and in some cases (like CVaR) have key theoretical properties.

Consider n risky assets. Let $\xi = (\xi_1, \ldots, \xi_n)^{\mathrm{T}} \in \mathbb{R}^n$ denote the uncertain returns of the n risky assets from the current time $t = 0$ to a fixed future time $t = 1$. Let $x = (x_1, \ldots, x_n)^{\mathrm{T}}$ denote a portfolio on these assets.

A (single-period) risk-reward problem aims at finding the portfolio x to be constructed at $t = 0$, in order to minimize the the risk of the portfolio's return, subject to the portfolio having a given minimum expected return μ_o. Formally, the risk-reward problem can be written as the following optimization problem:

$$\begin{aligned} \min \quad & \mathcal{R}(x^{\mathrm{T}}\xi) \\ \text{s.t.} \quad & \mathbb{E}(x^{\mathrm{T}}\xi) \geq \mu_0 \\ & x \in \mathcal{X} \end{aligned} \tag{1.1}$$

where $\mathbb{E}(\cdot)$ denotes expectation, $\mathcal{X}$ is the set of constraints for the assets in portfolio construction, for example: $\mathcal{X} = \{\sum_{i=1}^{n} x_i = 1, x \geq 0\}$ and $\mathcal{R}$ is some measure of risk like variance, semivariance, VaR or CVaR.

Model (1.1) is a generalization of Markowitz' classical concept of mean-variance optimization for the case of an arbitrary risk measure $\mathcal{R}$.

The problem's (1.1) computational complexity highly depends on the selected risk measure, the liquidity of the assets, and the constraints on the assets.

In particular, the problems in portfolio optimization are integer programs in the following cases,

1. *Liquidity of the assets*: The assets may be real world projects, which are non-divisible assets. You either choose to invest to the project or leave it. Thus, the decision to invest or not in these projects can be modeled with binary variables, which makes problem (1.1) to be an integer problem. Portfolio optimization with non-divisible assets are studied in in Chapter 2.

2. *Risk Measures*: Modeling risk-reward portfolio optimization problem with some particular risk measures may bring the integrality to the problem. For instance, if VaR is chosen as the risk measure, then one can rely on order estimators to formulate the problem as an Integer Program (IP). The VaR portfolio optimization problem is studied in Chapter 3.

3. *Trading Constraints*: These requirements come from real-world trading practice. Risk-reward portfolio optimization models typically results portfolios with

a large number of assets having small holdings. This is not very desirable because of the transaction costs of the assets, minimum lot sizes, management complexity and policies of the companies. To avoid these undesired portfolios, one can limit the number of assets in the portfolio by using cardinality constraints, and/or threshold constraints can be used prescribe lower and upper bounds on the fraction of capital invested in each asset and so on. Although trading constraints are not considered directly in this study, they can be adapted to the problems studied in Chapters 2 & 3 and some of the techniques developed in this chapters can be used to solve these modified problems.

In Chapter 4, we study a frequently encountered risk management problem in industrial gas networks. In this problem, the uncertainty in flow sensor measurements propagates and results as the uncertainty in system's output. System output is often associated with system's total cost, and the uncertainty in the cost is not plausible and associated with risk from the system manager's perspective. To detect, identify, and determine the sensor faults, we develop a systematic approach based on outlier detection, bias detection and noise dressing. The proposed methodology is applied to a simplified gas network model, where the corresponding results can be readily validated. Later, we apply the methodology to a real industrial gas network in the US. The analysis and results have shown us that prioritizing the sensor improvements can help practitioners to identify the key sensors subject to a budget to decrease the total system cost and it's variability throughout the time.

Finally, we investigate the out of sample performance of the application of hierarchical clustering algorithms in portfolio allocation strategies. The modern portfolio theory relies on estimation of expected returns and risks, but even small errors on these estimations can result in large deviations from optimal allocations. Thus, the predicted and realized risk returns often times deviate significantly from each other, which is not to the benefit of a risk-averse investor. Recent studies by Tola et al.

(2008), Pantaleo et al. (2011) and López de Prado (2016) have shown the benefits of incorporating clustering approach into portfolio allocation strategies. However, to the best of our knowledge there is no empirical study comparing some of these methodologies with each other. In his famous study, (DeMiguel et al. 2009) evaluates the out-of-sample performance of the sample-based mean-variance model, and its extensions designed to reduce estimation error, relative to the naive 1/N portfolio. By using the experimental design and some of the comparison measures from this study, we replicate some of those experiments to evaluate the out of sample performances of different clustering based portfolio allocation strategies and compare their performance with the classical portfolio selections strategies performance across different datasets. The preliminary results indicate that the integration of clustering to mean-variance models and its' extensions do not affect the out-of-sample performance of these models significantly when they are short-sales constrained. Similarly, the other clustering based portfolio allocation rules investigated in this study do not outperform the classical approaches consistently.

Chapter 2

Linear solution schemes for mean-semivariance project portfolio selection problems: An application in the oil and gas industry

2.1 Project Portfolio Selection in Oil and Gas Industry

A keystone economic sector where the problem of selecting an appropriate portfolio of project investments arises is the upstream oil and gas industry. In this sector, the project investment's returns are subject to high uncertainty, mainly driven by factors like geology, equipment costs, oil selling price, well production levels, and oil quality, among others. In a typical project, the profit's probability distribution is

usually asymmetrical (skewed), exhibiting a high probability of low profits and a low probability of high profits (Walls 2004). Moreover, given the significant amount of investment required to carry out a project, managers and investors in this industry have a strong bias against underperforming portfolios (Merritt et al. 2000, Quintino et al. 2013, Suslick and Schiozer 2004, Tyler et al. 2001), leaning towards downside-risk measures to quantify the risk of investment (Sira 2006).

Although different downside-risk measures are available in the literature (cf., Boasson et al. 2011, Jarrow and Zhao 2006, Markowitz et al. 1993, Rockafellar and Uryasev 2000), in this paper we focus on the semivariance risk measure. Through this measure, projects with a high probability of having returns lower than a critical value (e.g., the expected value or any other value specified by the decision maker) are considered risky. In other words, the semivariance does not consider values beyond the critical value (i.e., gains) as risk; thus, it is a more appropriate measure when investors are worried about portfolio underperformance (Markowitz et al. 1993).

The semivariance is a widely used measure of risk in the oil and gas industry. For example, Orman et al. (1999) propose an optimization routine in which the portfolio's semistandard deviation (square root of the semivariance) is minimized, subject to budget constraints and a target value for the expected Net Present Value (NPV). By varying this target, the authors construct an efficient frontier. Then, they find the optimal investment selection for each project based on a predetermined set of projects. In a more recent work, Sira (2006) uses scatter search to heuristically approximate an efficient portfolio frontier in the petroleum industry. This approach is used to determine how much investment must be allocated in a fixed set of projects. After comparing portfolios that minimize both variance and semivariance of the project portfolio's return, the author argues that the latter is preferable as a measure of risk in petroleum projects. Similar to Sira (2006), we consider the problem of finding a portfolio of projects; that is, *non-divisible* assets with minimum semivariance, but

where the projects to be included in the portfolio, rather than fixed, can be selected from a set of available investment projects.

To address this problem we first consider the more common portfolio allocation problem where the portfolio assets are divisible. In his seminal work on risk-reward portfolio selection, Markowitz (1952) proposed the use of the portfolio returns' variance as a measure of risk, and developed an optimization problem, together with a solution method, to obtain the portfolio selection that has minimum risk among those with a required expected return. This problem is now commonly referred as the Mean-Variance (MV) portfolio selection problem. Similar to the classical MV problem, Markowitz et al. (1993) proposed a quadratic programming formulation for the Mean-SemiVariance (MSV) portfolio selection problem, which is obtained using a *sampling approach* to estimate the problem parameters; that is, an estimation of the asset return distributions is obtained from a finite number of samples. These samples are typically obtained from historical data, simulations, or a combination of both. Thus, these portfolio selection problems have the characteristic that no specific distributional assumption about the asset return distributions is required to formulate or solve the corresponding selection problem.

The Mean-SemiVariance Project (MSVP) portfolio selection problem, a MSV problem with non-divisible assets, can be formulated as a Mixed-Integer Quadratic Programming (MIQP) problem for which specialized MIQP solvers can be used. However, unlike the MV problem formulation whose size only depends on the number of assets, the size of the MSVP problem formulation grows with both the number of non-divisible assets, and the number of samples used to estimate the problem's parameters, thus leaving open some concerns regarding scalability and solvability of the MSVP problem via MIQP solvers. Although existing solution methods for Quadratic Programming (QP) are quite competitive, the introduction of integer variables significantly increases the complexity of solving a MIQP problem and limits the size of

the problems that can be solved (Mansini et al. 2013). Similar challenges have been addressed for MV problems with integer variables (due to, e.g., transaction costs, cardinality constraints, lot size) by proposing solution approaches that avoid using MIQP solvers (cf., Lejeune 2013, Bertsimas and Shioda 2009).

To tackle the inherent difficulty in solving the MSVP problem, we propose two *linear* solution schemes that avoid the use of QP methods and only require the use of Mixed-Integer Linear Programmming (MILP) techniques. These approaches are useful alternatives to the MIQP when either because of problem size, solution time requirements, software requirements, or expertise, it is not suitable to directly use a MIQP solver. The first scheme is obtained from a natural approximation of the portfolio's semivariance that can be reformulated as a MILP problem. This MILP approximation is (formally) shown to work as an accurate proxy of the MSVP problem when the projects' NPVs are positively correlated, which is the case in our oil and gas industry problem. Furthermore, we develop a second linear solution scheme that requires the solution of a series of MILP problems for general instances of the MSVP problem. This scheme works even in the case of NPVs having arbitrary correlations (i.e., not all are positively correlated).

The proposed schemes have both practical and computational advantages compared to MIQP formulation. They might be more suitable for practitioners that are well acquainted with MILP techniques (Bixby 2002), but not with more advanced MIQP techniques. Also, the software required to solve the corresponding MIQP may require an additional investment over regular software required to solve MILP problems. More importantly, both solution schemes have the ability to solve instances of the MSVP problem that might not be possible to solve efficiently using MIQP solvers. Our linear solution schemes also contribute to the rich literature on using linear methods for portfolio allocation problems (see Mansini et al. 2013, for a recent review).

The remainder of the article is organized as follows. In Section 2.2, we formally introduce the MSVP problem. In Section 2.3.1, we present a linear approximation of the MSVP problem that requires the solution of a single MILP problem. Also, we quantitatively characterize the MILP approximation's effectiveness. In Section 2.3.2, we present a linear solution scheme capable of solving general MSVP instances by iteratively solving a series of MILP problems. Our computational results are presented in Section 2.4, where we illustrate the effectiveness of the linear solution schemes by solving a MSVP problem arising in a Latin American oil and gas company. In Section 2.5 we solve general instances of the MSVP problem that are constructed using data from the PSPLIB library of project scheduling problems (Kolisch and Sprecher 1997). In Section 2.6, we conclude the chapter with some final remarks.

2.2 Mean-semivariance project portfolio selection problem

In this section we formally introduce the MSVP problem. Consider n risky *non-divisble* investment projects. Let $r = (r_1, \ldots, r_n)^{\mathrm{T}} \in \mathbb{R}^n$ denote the uncertain NPVs of the n risky projects, which are calculated over a time horizon of T periods. Let $x = (x_1, \ldots, x_n)^{\mathrm{T}} \in \{0,1\}^n$ denote a portfolio on these projects; that is, the binary variables x_i take the value of 1 if the company invest in project i and 0 otherwise, for $i = 1, \ldots, n$. Thus, the portfolio's NPV is given by

$$r^{\mathrm{T}}x = x^{\mathrm{T}}r = \sum_{i=1}^{n} x_i r_i.$$

A (single-period) MSVP problem aims at finding the portfolio of projects $x \in \{0,1\}^n$ at time $t = 0$ that minimizes the semivariance of the project portfolio's NPV, subject to a given minimum expected NPV. Formally, the MSVP problem can be

written as the following optimization problem:

$$\begin{aligned} \min \quad & \mathbb{E}(\min\{0, x^{\mathrm{T}} r - \mathbb{E}(x^{\mathrm{T}} r)\}^2) \\ \text{s.t.} \quad & \mathbb{E}(x^{\mathrm{T}} r) \geq \mu_0 \\ & x \in \mathcal{X} \cap \{0,1\}^n, \end{aligned} \tag{2.1}$$

where $\mathbb{E}(\cdot)$ denotes expectation; $\mu_0 \in \mathbb{R}$ is the given minimum expected portfolio NPV; and $\mathcal{X} \subseteq \mathbb{R}^n$ is a given set defined by linear constraints, which might be used to enforce some relevant business conditions such as a budget constraint (*i.e.*, $\sum_{i=1}^{n} c_i x_i \leq B$, where c_i is the investment required for i-th project and B is the total available budget). For the MSVP problem in the oil and gas industry considered here, a detailed description of the set $\mathcal{X}$ is provided in Section 2.4. Here, we choose 0 as the *critical value* (e.g., Markowitz et al. 1993) to define the downside semivariance. However, our results extend in straightforward fashion for other choices of the critical value, such as a market benchmark (cf., Markowitz et al. 1993).

It is clear from (2.1) that the MSVP problem is analogous to a classical risk-reward portfolio allocation problem with illiquid assets in which the risk is measured by the portfolio returns' semivariance, and the reward is the expected portfolio's return.

In order to solve (2.1), we use a *sampling approach* (cf., Birge and Louveaux 2011, Konno and Yamazaki 1991, Lejeune 2013, Markowitz et al. 1993, Rockafellar and Uryasev 2000), in which an estimation of the distribution of the random variables of interest is obtained from a finite number of samples $r^1, \ldots, r^m \in \mathbb{R}^n$. These samples are typically obtained from historical data, simulations, or a combination of both. Using this sampling approach, the MSVP problem in (2.1) can be written as:

$$\begin{aligned} \min \quad & \frac{1}{m}\sum_{j=1}^{m} \min\{0, x^{\mathrm{T}} r^j - x^{\mathrm{T}}\mu\}^2 \\ \text{s.t.} \quad & x^{\mathrm{T}}\mu \geq \mu_0 \\ & x \in \mathcal{X} \cap \{0,1\}^n, \end{aligned} \tag{2.2}$$

where the vector $\mu = (\mu_1, \ldots, \mu_n)^{\mathrm{T}} \in \mathbb{R}^n$ of mean project return estimates is obtained by letting

$$\mu_i = \frac{1}{m}\sum_{j=1}^{m} r_i^j, \tag{2.3}$$

for $i = 1, \ldots, n$.

For ease of exposition, we will use (2.3) to obtain $\mu \in \mathbb{R}^n$ in our numerical experiments; however, our results are independent of this choice, and a variety of other estimation methods can be used. Also, note that to obtain an asymptotically unbiased and strongly consistent estimator of the semivariance, we should use the factor $\frac{m}{(m-1)^2}$ instead of $\frac{1}{m}$ in the objective function of (2.2) (Josephy and Aczel 1993). However, for the sake of clarity, we will use the latter, as changing this factor does not affect the composition of the optimal project selection.

After introducing the auxiliary variable y_j, which captures the value $\min\{0, x^{\mathrm{T}} r^j - x^{\mathrm{T}}\mu\}$ for each $j = 1, \ldots, m$, the MSVP problem in (2.2) can be written as an optimization problem with a convex quadratic objective, linear constraints, and binary variables. This result is formalized in Proposition 1.

Proposition 1 (Markowitz et al. (1993)) *The mean-semivariance project port-*

folio selection problem (2.2) *is equivalent to:*

$$\begin{aligned} \min \quad & \frac{1}{m}\sum_{j=1}^{m} y_j^2 \\ \text{s.t.} \quad & y_j \leq x^{\mathrm{T}}(r^j - \mu) \quad j = 1, \ldots, m \\ & y_j \leq 0 \qquad\qquad\quad j = 1, \ldots, m \\ & x^{\mathrm{T}}\mu \geq \mu_0 \\ & x \in \mathcal{X} \cap \{0,1\}^n. \end{aligned} \tag{2.4}$$

Furthermore, the objective function in (2.4) *is convex.*

Proposition 1 shows that the MSVP problem is a MIQP problem, that is, an optimization problem with a convex quadratic objective and linear constraints, with the additional constraint of some of its variables being integer (more specifically, in (2.4) variables are required to be binary). Thus, the MSVP formulation in (2.4) can be solved using branch-and-bound (cf., Borchers and Mitchell 1994, Nemhauser and Wolsey 1988) in conjunction with QP techniques (cf., Bienstock 1996). In particular, `CPLEX`, `Gurobi`, and `Xpress-MP` are among the commercial optimization solvers that offer special solution algorithms for MIQP problems based on such techniques.

2.3 Linear solution schemes

In this section we show that the MSVP problem in (2.4) can be efficiently solved without using QP solvers; that is, it can be solved using branch-and-bound in conjunction with linear programming techniques. We refer to these solution methodologies as *linear* solution schemes. Besides substantially enlarging the number of optimization solvers that can be used to solve the MSVP problem, these linear solution schemes allow us to solve large-scale instances of the MSVP problem much faster than with a MIQP approach.

Note that the MIQP problem in (2.4) can easily become a large-scale problem when either the number of projects, n, or the number of samples used to estimate the distribution of the projects' NPVs, m, is large. Clearly, this behavior results from n and m being the dimension of the x- and y-variables in (2.4), respectively. In this regard, we emphasize the difference in the project portfolio selection problem when the variance is used as a measure of risk. As opposed to semivariance, using the variance implies the solution of a single MIQP problem whose size depends on the number of candidate projects, but not on the number of samples used to estimate the mean and the variance-covariance matrix of project's NPVs. Further, in order to solve a single MIQP problem, it is necessary (loosely speaking) to solve a large number of (potentially large) QP problems (relaxed MIQP problems), obtained by branching on the corresponding binary variables.

For the reasons discussed above, in Section 2.3.1 we first introduce a MILP formulation that accurately approximates the solution of the MSVP problem when the projects' NPVs are positively correlated and the total number of projects is moderate. Next, in Section 2.3.2 we show that a general class of the MSVP problem, and in particular instances of the problem with a large number of projects and samples, can be solved efficiently by solving a sequence of MILP problems using a *Benders decomposition* approach in which the *Benders cuts* (cf. Nemhauser and Wolsey (1988, Sections II.3.7 and II.5.4), and Freund (2004)) are computed in closed-form.

2.3.1 MILP approximation for MSVP portfolio selection problem

In this section we present an approximation for the MSVP problem in (2.4), which is obtained by solving a *single* MILP problem with as many binary variables as the corresponding MIQP. We begin by stating the following optimization problem related

to (2.2):

$$\begin{aligned} \min \quad & \frac{1}{m}\sum_{i=1}^{n}\sum_{k=1}^{n}\widetilde{\sigma}_{ik}x_i x_k \\ \text{s.t.} \quad & x^{\mathrm{T}}\mu \geq \mu_0 \\ & x \in \mathcal{X} \cap \{0,1\}^n, \end{aligned} \tag{2.5}$$

where

$$\widetilde{\sigma}_{ik} = \sum_{j=1}^{m} \min\{0, r_i^j - \mu_i\} \min\{0, r_k^j - \mu_k\}. \tag{2.6}$$

for $i = 1, \ldots, n$, $k = 1, \ldots, n$. First, we will show that (2.5) is a pessimistic approximation to (2.2); that is, (2.5) overestimates the semivariance of the project's portfolio in (2.2), making (2.5) potentially more suitable for risk-averse investors. Then, we will show that the more positively correlated the projects in the portfolio are, the better (2.5) works as an approximation to (2.2). Even though this condition seems overly restrictive, there is strong evidence that positive correlations are ubiquitous in the oil and gas industry, in part, because most projects are influenced by the same economic and market conditions (e.g. interest rates, oil prices, and gas prices). Further evidence of this will be given in Section 2.4. Finally, we will show that (2.5) can be rewritten as a MILP problem by introducing appropriate extra continuous variables.

To see that (2.5) provides a pessimistic approximation to (2.2), let $u \in \mathbb{R}^n$ be given, and define $I^- = \{i : u_i < 0, i = 1, \ldots, n\}$, and $I^+ = \{i : u_i \geq 0, i = 1, \ldots, n\}$. Clearly,

$$0 \geq \min\left\{0, \sum_{i=1}^{n} u_i\right\} = \begin{cases} \displaystyle\sum_{i\in I^-} u_i + \sum_{i\in I^+} u_i, & \text{if } \left|\displaystyle\sum_{i\in I^-} u_i\right| \geq \left|\displaystyle\sum_{i\in I^+} u_i\right|; \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

Also

$$\sum_{i=1}^{n} \min\{0, u_i\} = 0 + \sum_{i \in I^-} u_i. \tag{2.8}$$

Using (2.7) and (2.8) in the two cases $\left|\sum_{i\in I^-} u_i\right| \geq \left|\sum_{i\in I^+} u_i\right|$ and $\left|\sum_{i\in I^-} u_i\right| \leq \left|\sum_{i\in I^+} u_i\right|$, it follows that

$$0 \geq \min\left\{0, \sum_{i=1}^{n} u_i\right\}, \text{ and } \min\left\{0, \sum_{i=1}^{n} u_i\right\} \geq \sum_{i=1}^{n} \min\{0, u_i\}, \tag{2.9}$$

and therefore:

$$\left(\min\left\{0, \sum_{i=1}^{n} u_i\right\}\right)^2 \leq \left(\sum_{i=1}^{n} \min\{0, u_i\}\right)^2. \tag{2.10}$$

With (2.10), and letting $\widetilde{r}^j := r^j - \mu$, $j = 1, \ldots, m$, we have that the objective function of (2.2) can be bounded from above as follows:

$$\begin{aligned} \frac{1}{m}\sum_{j=1}^{m} \min\left\{0, \sum_{i=1}^{n} x_i \widetilde{r}_i^j\right\}^2 \; & \leq \frac{1}{m}\sum_{j=1}^{m}\left(\sum_{i=1}^{n} \min\{0, x_i\widetilde{r}_i^j\}\right)^2 = \\ & \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{k=1}^{n} \min\{0, x_i\widetilde{r}_i^j\}\min\{0, x_k\widetilde{r}_k^j\}) = \\ & \frac{1}{m}\sum_{i=1}^{n}\sum_{k=1}^{n} x_i x_k \left(\sum_{j=1}^{m} \min\{0, \widetilde{r}_i^j\}\min\{0, \widetilde{r}_k^j\}\right) = \\ & \frac{1}{m}\sum_{i=1}^{n}\sum_{k=1}^{n} \widetilde{\sigma}_{ik} x_i x_k. \end{aligned} \tag{2.11}$$

The first inequality follows from (2.10), and the second to last equality follows from $x_i \geq 0$, $i = 1, \ldots, n$. Hence, (2.5) is a pessimistic approximation to (2.2) because its objective overestimates the expected squared downside deviations; that is, the semivariance.

Notice that (2.5) will be equivalent to (2.2) whenever the second inequality in (2.9) holds with equality when replacing $u = x^{\mathrm{T}}\widetilde{r}^j$, for all $j = 1, \ldots, m$. The second inequality in (2.9) holds with equality when $I^- = \{1, \ldots, n\}$ or $I^+ = \{1, \ldots, n\}$. That is, problems (2.5) and (2.2) will be equivalent if $I^{j+} := \{i \in \{1, \ldots, n\} : x_i\widetilde{r}_i^j \geq 0, \} =$

$\{1,\ldots,n\}$ or $I^{j^-} := \{i \in \{1,\ldots,n\} : x_i\widetilde{r}_i^j < 0\} = \{1,\ldots,n\}$, for all $j = 1,\ldots,,m$. Clearly, for all the samples to satisfy that the deviations in a sample be either all above the mean or all below the mean, the NPVs of each project must be highly correlated. As discussed in Section 2.4, for MSVP problems arising in the oil and gas industry, it is reasonable to expect (real-world) scenarios with high correlations where this approximation works remarkably well.

The objective function of (2.5) can be linearized by introducing appropriate extra continuous variables. Let $I_\sigma^+ := \{(i,k) : \widetilde{\sigma}_{ik} > 0, i = 1,\ldots,n, k = 1,\ldots,n\}$, and $I_\sigma^- := \{(i,k) : \widetilde{\sigma}_{ik} \leq 0, i = 1,\ldots,n, k = 1,\ldots,n\}$. Then problem (2.5) is equivalent to the following MILP problem:

$$\begin{array}{lll} \min & \dfrac{1}{m}\displaystyle\sum_{(i,k)\in I_\sigma^+} \widetilde{\sigma}_{ik} y_{ik} & \\ \text{s.t.} & x^{\mathrm{T}}\mu \geq \mu_0 & \\ & y_{ik} \geq x_i + x_k - 1 & \text{for all } (i,k) \in I_\sigma^+ \\ & y_{ik} \geq 0 & \text{for all } (i,k) \in I_\sigma^+ \\ & x \in \mathcal{X} \cap \{0,1\}^n. & \end{array} \tag{2.12}$$

The equivalence between (2.5) and (2.12) follows from the next observations. First, from (2.6) it follows that $I_\sigma^- = \{(i,k) : \widetilde{\sigma}_{ik} = 0, i = 1,\ldots,n, k = 1,\ldots,n\}$. Second, if $(i,k) \in I_\sigma^+$, then $y_{ik} \geq x_i + x_k - 1$ and $y_{ik} \geq 0$ imply that $y_{ik} \geq x_i x_k$, but since $\widetilde{\sigma}_{ik} > 0$, then at any optimal solution of (2.12), y_{ik} would be at its lower bound $y_{ik} = x_i x_k$.

2.3.2 Benders-based linear solution scheme for MSVP portfolio selection problems

In this section, we present a linear solution scheme for the MSVP problem that is based on a suitable use of the *Benders decomposition* technique (cf. Nemhauser and Wolsey (1988, Sections II.3.7 and II.5.4), and Freund (2004)). To make the

presentation more succinct, we re-state (2.4) as follows:

$$
\begin{aligned}
\min \quad & \frac{1}{m} y^{\mathrm{T}}(I) y \\
\text{s.t.} \quad & y \leq \widetilde{R} x \\
& y \leq 0 \\
& x \in \mathcal{X}' \cap \{0,1\}^n,
\end{aligned}
\tag{$\mathcal{S}$}
$$

where $y := [y_j]_{j=1,\ldots,m}$, I is the $m \times m$ identity matrix, $\widetilde{R}$ is a $m \times n$ matrix, whose row j is given by $[\widetilde{R}]_j := (r^j - \mu)^{\mathrm{T}}$, $j = 1, \ldots, m$, and $\mathcal{X}' := \mathcal{X} \cup \{x \in \mathbb{R}^n : x^{\mathrm{T}} \mu \geq \mu_0\}$.

The idea of a Benders decomposition approach is to divide the problem variables into two groups: the *complicating* and the *non-complicating* variables. One begins by fixing the complicating variables in the original problem to a particular value. The resulting problem –so-called *Benders subproblem*– should be solvable to optimality, and in particular, the *dual* (see, e.g. Fang and Puthenpura (1993, Chapter 9.1.2)) of the Benders subproblem should be solvable to optimality. The dual solution of the Benders subproblem is then used to construct a *Benders master problem* on the complicating variables of the original problem. Solving iteratively both the Benders subproblem and master problem leads to a solution of the original problem that might be obtained faster than by solving the (full) original problem. For the MSVP problem, next we show that with an appropriate choice of the complicating variables, the Benders subproblem can be solved in closed-form.

To address problem ($\mathcal{S}$) via a Benders decomposition approach, we choose the x variables as the complicating variables in ($\mathcal{S}$). After fixing the x variables to a value $\hat{x} \in \mathcal{X}' \cap \{0,1\}$, and (for convenience) making the change of variable $y \to -y$, we

obtain the problem:

$$\begin{array}{lll} \min & \dfrac{1}{m} y^{\mathrm{T}}(I) y & \\ \text{s.t.} & y \geq -\widetilde{R}\hat{x} & (u) \\ & y \geq 0 & (u_0), \end{array} \tag{2.13}$$

where $u \in \mathbb{R}^m$ are the dual variables associated to the return constraints and $u_0 \in \mathbb{R}^m$ are the dual variables associated to the non-negativity constraints in (2.13). The (convex) quadratic program in (2.13) corresponds to the Benders subproblem, whose *Wolfe dual* is given by (see, e.g., Nocedal and Wright (2006, Chapter 12)):

$$\begin{array}{ll} \max & -u^{\mathrm{T}}\widetilde{R}\hat{x} - \dfrac{1}{m} y^{\mathrm{T}}(I) y \\ \text{s.t.} & -\dfrac{2}{m} y + u + u_0 = 0 \\ & u, u_0 \geq 0, \end{array} \tag{2.14}$$

Problem (2.14) is equivalent to:

$$\begin{array}{ll} \max & -u^{\mathrm{T}}\widetilde{R}\hat{x} - \dfrac{m}{4}(u+u_0)^{\mathrm{T}}(u+u_0) \\ \text{s.t.} & u, u_0 \geq 0. \end{array} \tag{2.15}$$

In any optimal solution of (2.15) we have $u_0 = \vec{0}$, so (2.15) is equivalent to:

$$\begin{array}{ll} \max & \displaystyle\sum_{j=1}^{m} \left(-(\hat{x}^{\mathrm{T}}\widetilde{r}^j) u_j - \frac{m}{4} u_j^2 \right) \\ \text{s.t.} & u_j \geq 0, j = 1, \ldots, m. \end{array} \tag{2.16}$$

Notice that problem (2.16) decomposes into m independent problems:

$$\begin{array}{ll} \max & -(\hat{x}^{\mathrm{T}}\widetilde{r}^j) u_j - \dfrac{m}{4} u_j^2 \\ \text{s.t.} & u_j \geq 0, \end{array} \tag{2.17}$$

for $j = 1, \ldots, m$; which can be solved by inspection: If $(\hat{x}^{\mathrm{T}}\widetilde{r}^{j}) \geq 0$, then the optimal solution of (2.17) is $u_j^* = 0$. If $(\hat{x}^{\mathrm{T}}\widetilde{r}^{j}) < 0$, then we get a concave quadratic objective in (2.17):

$$|(\hat{x}^{\mathrm{T}}\widetilde{r}^{j})|u_j - \frac{m}{4}u_j^2$$

that has a maximum at $u_j^* = \frac{2}{m}|(\hat{x}^{\mathrm{T}}\widetilde{r}^{j})|$. So the optimal solution $u^*(\hat{x}) \in \mathbb{R}^m$ of the Benders dual subproblem (2.14) can be obtained in closed-form as follows:

$$u_j^*(\hat{x}) = \begin{cases} 0 & \text{if } (\hat{x}^{\mathrm{T}}\widetilde{r}^{j}) \geq 0, \\ \frac{2}{m}|(\hat{x}^{\mathrm{T}}\widetilde{r}^{j})| & \text{if } (\hat{x}^{\mathrm{T}}\widetilde{r}^{j}) < 0, \end{cases} \tag{2.18}$$

for $j = 1, \ldots, m$. With the Benders dual subproblem solution, the Benders master problem is constructed as follows. Given a finite index set $\mathcal{K}$, and a set of feasible portfolios $\hat{\mathcal{X}}'_{\mathcal{K}} = \{\hat{x}_k \in \mathcal{X}' \cap \{0,1\}^n : k \in \mathcal{K}\}$, consider the Benders master problem

$$\begin{aligned} \min \quad & q \\ \text{s.t.} \quad & q \geq \sum_{j=1}^{m} -(x^{\mathrm{T}}\widetilde{r}^{j})u_j^*(\hat{x}_k) - \frac{m}{4}u_j^*(\hat{x}_k)^2; \quad \forall \hat{x}_k \in \hat{\mathcal{X}}'_{\mathcal{K}} \\ & x \in \mathcal{X}' \cap \{0,1\}^n. \end{aligned} \tag{$\mathcal{M}(\hat{\mathcal{X}}'_{\mathcal{K}})$}$$

Note that the right-hand side of the first set of constraints in $(\mathcal{M}(\hat{\mathcal{X}}'_{\mathcal{K}}))$ is closely related to the objective function of the Benders dual subproblem (2.15).

With a closed-form expression for the solution of the Benders dual subproblem, and with the construction of the Benders master subproblem given in $(\mathcal{M}(\hat{\mathcal{X}}'_{\mathcal{K}}))$, we can now state in Algorithm 1, a Benders-based solution algorithm for the MSVP problem.

After execution, Algorithm 1 returns an ϵ-optimal portfolio solution x_ϵ^*. That is,

Algorithm 1 Benders linear solution scheme for the MSVP problem

1: **procedure** MSVP_BENDERS($\epsilon > 0$)
2: $\mathcal{K} \leftarrow \emptyset$, $k = 1$, GAP$= \infty$
3: **while** GAP$> \epsilon$ **do**
4: **compute** $\hat{x}_k$, z_k, the optimal solution and objective of $(\mathcal{M}(\hat{\mathcal{X}}'_{\mathcal{K}}))$
5: **compute** $u^*(\hat{x}_k)$ using (2.18)
6: $\mathcal{K} \leftarrow \mathcal{K} \cup k$, $k \leftarrow= k + 1$
7: UPPBOUND $\leftarrow \sum_{j=1}^{m} -(\hat{x}_k^{\mathrm{T}} \widetilde{r}^j) u_j^*(\hat{x}_k) - \frac{m}{4} u_j^*(\hat{x}_k)^2$, LOWBOUND$_k \leftarrow z_k$
8: GAP$\leftarrow$ |UPPBOUND $-$ LOWBOUND$_k$|/|LOWBOUND$_k$|
9: **end while**
10: **return** $x_\epsilon^* = \hat{x}_k$
11: **end procedure**

if we let $x^* := \mathrm{argmin}_x\{\mathcal{S}\}$ be the optimal mean-semivariance project portfolio, then

$$\frac{\mathrm{SV}(x_\epsilon^*) - \mathrm{SV}(x^*)}{\mathrm{SV}(x^*)} < \epsilon, \tag{2.19}$$

where SV represents semivariance of any portfolio of projects $x \in \{0,1\}^n$, given by

$$\mathrm{SV}(x) := \frac{1}{m} \sum_{j=1}^{m} \min\{0, x^{\mathrm{T}} r^j - x^{\mathrm{T}} \mu\}^2.$$

The proof of convergence for Algorithm 1 follows from (Flippo and Rinnooy-Kan 1993).

Note that the Benders-based linear solution scheme for the MSVP problem outlined in this section requires, at its core, the iterative solution of MILPs in `Step 4` of the algorithm. This is because the non-linearity of the original problem's objective is handled in closed-form in `Step 5` of the algorithm. It is worth to mention that a *regularized* version (cf., Ruszczyński 1997) of the Benders-based algorithm outlined here for the MSVP problem can be implemented without changing the complexity of the Master problem in $(\mathcal{M}(\hat{\mathcal{X}}'_{\mathcal{K}}))$. Namely, following Ruszczyński (1997), the objective function in $(\mathcal{M}(\hat{\mathcal{X}}'_{\mathcal{K}}))$ can be changed to $c(q,x) := q + \frac{1}{\sigma}\|x - \hat{x}_k\|^2$ with $\sigma > 0$. Moreover, taking advantage of the fact that both $x, \hat{x}_k \in \{0,1\}^n$ it follows that $c(q,x)$

is equivalent to the following linear function $c(q, x) = q + \frac{1}{\sigma}(x - 2x\hat{x}_k + \hat{x}_k)$. This means that the MSPV problem can be solved via a Benders decomposition approach where the regularized Benders Master problem remains a MILP and the Benders cuts are found in closed-form. Although experiments were carried out with this regularized version of the Benders algorithm, the performance difference with the classical Benders Algorithm 1 are not significant, and in Section 2.5, we report results using the non-regularized Benders Algorithm.

2.4 Case study: project selection in an upstream oil and gas company

In this section we consider an instance of the MSVP problem arising in the oil and gas industry. After giving a detailed description of the problem in Section 2.4.1, in Section 2.4.2 we report the computational results of the linear solution scheme presented in Section 2.3.1.

2.4.1 Data and detailed model

The case study is based on our experience with an upstream oil and gas company operating in Latin America, which is one of the top 40 largest companies in the world. We consider a division of the company with 27 non-divisible candidate projects for investment along a 30-year planning horizon with an available budget of US$ 100 million per year and expected production for the first year of at least 40,000 barrels. Besides the known capital investment requirements and the production and operational costs, the projects are subject to precedence relations. For example, the execution of some projects require the execution of other complementary projects.

The NPV calculation for each project involves deterministic elements like the capi-

tal investment requirements and the production and operational costs. It also involves more volatile and stochastic components, like the project's production level modeled by triangular distributions for pessimistic, moderate, and optimistic scenarios and the international trade petroleum price (WTI), forecasted by a mean-reversion model (cf., Dixit and Pindyck 1994). It should be emphasized that, according to Sira (2006), the uncertain production levels and the oil prices account for 80% of the NPV's volatility in a typical petroleum project (for literature on forecasting petroleum prices, see Al-Harthy (2007)). We use Monte Carlo simulation to model the uncertainties, considering a variance reduction technique known as common random numbers (cf., Law and Kelton 2000) to ensure that the same realizations for the key underlying random variable, namely the WTI price, were used to calculate the NPV for all projects. These values are used to construct the vector μ used in the expected return constraint in (2.2); that is, μ_i corresponds to the average NPV of project i, for $i = 1, \ldots, n$ where $n = 27$. In A.2, Table A.1 displays the average NPV of the projects when estimated with different sample sizes. In addition, Figure 2.1 shows the skewed nature of the NPV for a typical oil and gas project (i.e., low profits are more likely to occur than high profits). In this case, 1,000 NPV realizations are produced using Monte Carlo simulation. Due to confidentiality agreements, the average NPV for each project has been modified by adding a constant. However, although this shift affects the probability of loss and the mean return of the projects, it does not affect the deviations from the mean used to measure the risk of the project's portfolio.

The NPVs of the considered oil and gas projects are highly correlated, given that they belong to the same industry and are affected by the same market conditions. Figure 2.2 shows a histogram of the upper triangular portion of the correlation matrix (excluding the diagonal) where it is worth noting that more than 75% of the pairwise correlations are higher than 0.80, all correlations are positive, and only 8% of the correlations are less than 0.1. Although the calculated correlations appear to be

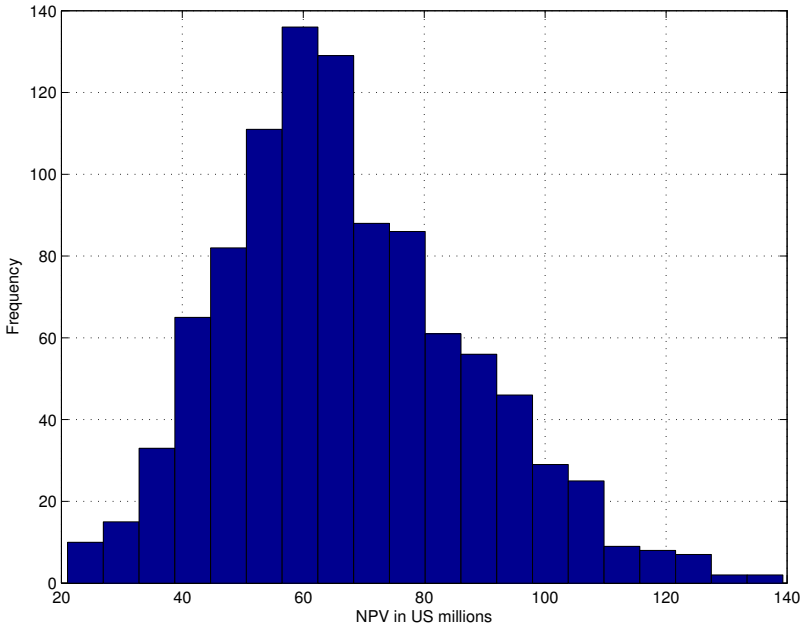


Figure 2.1: Histogram of 1,000 random realizations of the NPV (in US millions) of a typical oil and gas project obtained using Monte Carlo simulation.

overly high, evidence of positive and strong correlation between the projects in the same industry is ubiquitous in the literature. For instance, in Bodie et al. (2005), it is stated that correlations between security returns in the same industry tend to be positive because they are influenced by the same economic and market conditions. Thus, changes in economic factors such as interest rates, labor, and raw material cost affect simultaneously the performance of all companies and their projects in the same sector.

The linear constraints defining the set $\mathcal{X}$ in (2.1) for the oil and gas MSVP portfolio selection problem include a required minimum production level per period of the planning horizon; limiting budget constraints per time period; limits on the total production and operational cost per time period; and precedence relations between

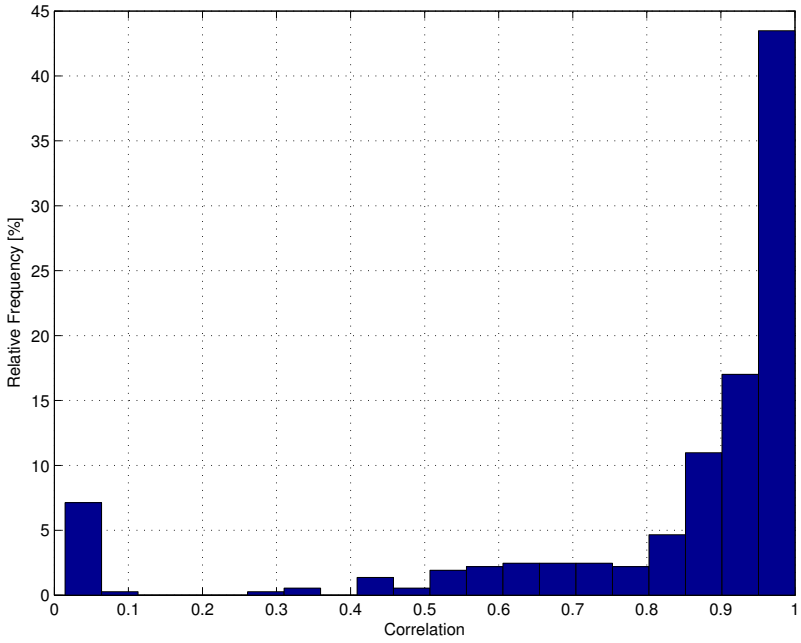


Figure 2.2: Histogram of the 27 project NPV correlations (excluding the diagonal) used in Table 2.1.

projects. Specifically, in this case we have

$$\mathcal{X} = \left\{ x \in \{0,1\}^n : \begin{array}{ll} \sum_{i=1}^{n} q_{it} x_i \geq w_t, & t \in T; \\ \sum_{i=1}^{n} k_{it} x_i \leq b_t, & t \in T; \\ \sum_{i=1}^{n} c_{it} x_i \leq h_t, & t \in T; \\ x_i \leq x_j, & (i,j) \in A \end{array} \right\}. \tag{2.20}$$

In 2.20, set T represents the time periods within the planning horizon. Parameters q_{it}, c_{it}, and k_{it} are the expected barrel production, the production and operational costs, and the capital investment requirements of project i in time period t, respectively. Parameters w_t, h_t, and b_t are the minimum production level, the maximum allowable production and operational costs, and the available budget for investment in period t, respectively. Note that, although variable x_i is not indexed in t, the time is implicitly

considered in the expected barrel production, the production and operational costs, and the capital investment requirements for each project per period of the planning horizon (i.e., parameters q_{it}, c_{it}, and k_{it}, respectively). That is, if project i is selected (i.e., variable x_i equals 1), its expected oil production and costs are accounted in the left-hand side of the constraints, in order to satisfy the minimum production level and the costs limits for each period of the planning horizon. Further, set A defines the precedence relations between projects; that is, if selecting project i implies the selection of project j, then $(i, j) \in A$. The complete list of precedence relations between the projects used in the case study is given in Eq. (A.1) in A.2.

Our algorithms are implemented in MATLAB and executed on a 64-bit workstation with AMD Opteron 2.0 GHz CPU and 32 GB of RAM. We use `CPLEX 12.5` to solve both the MILP approximation and the MIQP formulation to optimality.

2.4.2 Numerical results

In this section, computational experiments are conducted to show the accuracy and efficiency of the MILP approximation proposed in Section 2.3.1 to solve the oil and gas industry MSVP problem. Figure 2.3 displays the *semivariance efficient frontier* (i.e., plots the optimal project portfolio's semivariance for different values of μ_0) obtained after solving the MILP problem defined in (2.12) and the MIQP formulation in (2.4) with the side constraints $\mathcal{X}$ defined in (2.20). The number of projects and number of samples in the problems solved are $n = 27$ and $m = 1000$, respectively. Results in Figure 2.3 show that, thanks to the strong positive correlations of the projects in this case study, the MILP approximation effectively finds the set of non-dominated portfolios in the frontier. For practical purposes, this result implies that the MILP approximation in (2.12) can help decision makers to create a semivariance efficient frontier showing the tradeoff between risk and profitability, without the use of nonlinear programming techniques. The total time required to compute the efficient

frontier in Figure 2.3 using the MIQP approach is 84.04 s, whereas the total time required to compute it using the MILP approach is 20.79 s.

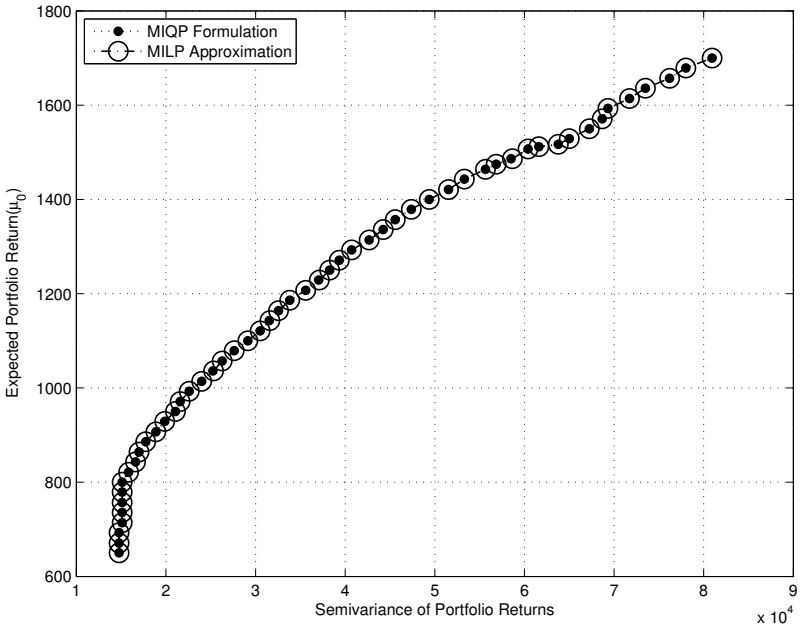


Figure 2.3: Semivariance efficient frontier of MSVP portfolios for fifty (50) different benchmark NPV return values, computed using the (exact) MIQP formulation (2.4) and the MILP approximation (2.12)

To further illustrate the performance of the MILP in (2.12) when the NVPs are highly correlated, we generate additional instances of the MSVP problem based on the original oil and gas data. Namely, we generate instances of $n = 27$ projects with sample sizes $m = 100$, 500, 1000, 3000, 5000, 7000, 9000, and 10000. To reach the desired value of m, additional samples are randomly drawn from the original data. Regardless of the sample size, we use $\mu_0 = 698$, as in the original oil and gas case study. We compare the proposed MILP approximation with the default `CPLEX 12.5` MIQP solver.

Table 2.1 shows the results obtained by the MILP and the MIQP models. The first column shows the sample size used to estimate the projects' NPVs return distributions. The resulting portfolio's semivariance (i.e., objective function values) are shown in the second and third columns, whereas the execution times are reported in

the fourth and fifth columns. The last column shows the computation time speedup, which is calculated as the ratio between the MIQP to the MILP execution times. The last row reports the geometric mean of the speedups for all the instances. In this case, we use the geometric mean because it avoids being overly optimistic with good ratios obtained on few instances (Lozano and Medaglia 2013).

The results summarized in Table 2.1 show that the MILP approximation finds the optimal portfolio's semivariance (i.e., the MIQP solution). As before, good accuracy performance of the MILP approximation is due to the positive correlated nature of the project's NPVs. Although the MIQP approach does slightly better than MILP approach in the instance with the smallest number of samples; overall, the geometric mean shows that the MILP approach is roughly eight times faster than the MIQP approach. This result is expected, given that the size of the MILP does not increase as the number of samples grows.

No.	**Portfolio's Semivariance**		**Time (s)**		
Samples	**MIQP**	**MILP Apx.**	**MIQP**	**MILP Apx.**	**Speedup**
100	14887	14887	0.47	0.75	0.62
500	15746	15746	1.21	0.48	2.54
1000	14769	14769	3.44	0.58	5.93
3000	14767	14767	4.79	0.60	7.94
5000	14768	14768	8.91	0.52	17.09
7000	14764	14764	15.23	0.55	27.48
9000	14769	14769	21.50	0.59	36.21
10000	14769	14769	34.66	1.53	22.66
Geo. Mean					8.55

Table 2.1: Computational results for instances of the MSVP problem based on the oil and gas case study with 27 projects, a minimum NPV benchmark return of 698, and number of samples ranging from 100 to 10000.

We further explore the quality of the MILP approximation scheme in (2.12) for the case when there are negative correlations present between the projects' NPVs. To do so, we use the same instance of Figure 2.3 and multiply the sample NPVs of thirteen

(13) projects (randomly selected) by -1. The results are shown in Figure 2.4, in which the quality of the MILP approximation decreases compared to the MIQP. However, even in this case the MILP approximation scheme provides a fair approximation of the MSVP efficient frontier. Note that due to this change on the instance data, the maximum expected NPV that can be obtained from the projects is now lower than in the original instance shown in Figure 2.3. In this case, the total time required to compute the efficient frontier in Figure 2.4 using the MIQP approach is 69.22 s, whereas the total time required to compute it using the MILP approach is 4.49 s.

Although it provides an accurate approximation to the semivariance when NPVs are positively correlated, the MILP scheme in (2.12) is limited by the fact that the number of continuous variables grows quadratically with the number of projects in the problem (i.e., as n^2).

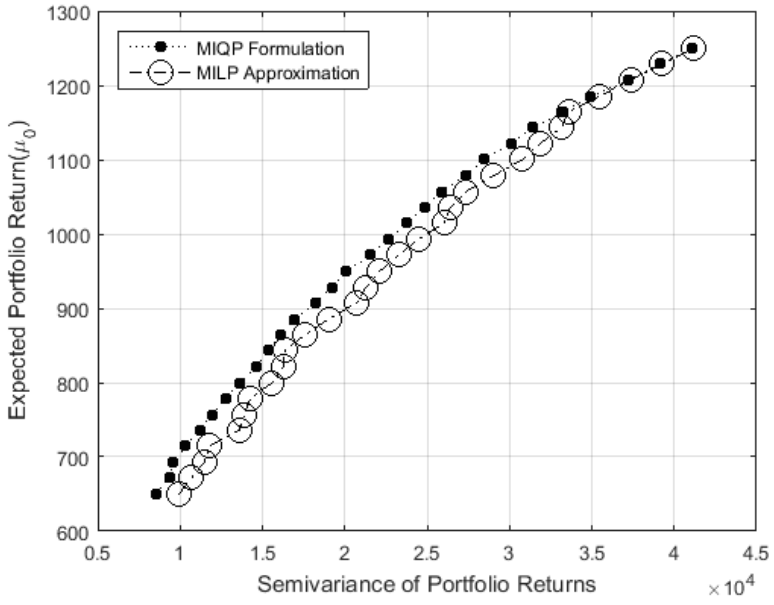


Figure 2.4: Semivariance efficient frontier of MSVP portfolios computed using the (exact) MIQP formulation and the MILP approximation (2.12) in when there are negative correlations present between the projects' NPVs.

2.5 General MSVP instances

In this section we study the accuracy and efficiency of the Benders-based solution approach proposed in Section 2.3.2 to solve general instances of the MSVP problem. . We test the limits of our approaches by considering instances with a large number of projects (n) and samples (m), and with multiple correlation levels among projects' NPVs. This analysis is motivated by the fact that some oil and gas companies may have a large number of candidate projects.

To see this, note that the case study considered in Section 4 arises from a project selection problem in one of the six divisions of an oil and gas company operating in Latin America. For the particular year of this analysis, the division's exploration budget was around US$100!M, which was 20% of the company's total exploration budget. To put the project selection problem of this division in context, in 2014 the top-ranked capital expenditures in exploration of some larger oil and gas companies ranged between US$1400 M and US$2500 M (EY 2015). Thus, from a budget perspective and considering the worldwide scale of operations of larger companies, MSVP problem instances with possibly hundreds of candidate projects may arise in practice. Additionally, the number of drilling permits approved by environmental authorities could be an estimate of the number of candidate projects in a company's portfolio. In 2014, the Oil and Gas Conservation Commission of the state of Wyoming alone approved 3786 different drilling permits, with some companies requesting permits for the exploration of more than 300 and up to 923 different wells (Oil and Commission 2015).

Also, the consideration of different correlation profiles among NPVs is motivated by the fact that current petroleum prices may encourage oil and gas companies to bring new types of projects into their portfolios (e.g., enhanced oil recovery, alternative refining processes, biofuels), which can be less (or even negatively) correlated

with the traditional exploration and production projects. Additionally, the datasets used in this section include realistic features arising in the oil and gas industry such as resource and precedence constraints, as well as skewed NPV's distributions. Section 2.5.1 describes the dataset generation procedure used to test our algorithms and Section 2.5.2 presents the computational results of the Benders-based solution approach described in Section 2.3.2.

2.5.1 Data

Given the absence of datasets for the MSVP problem in the literature, we generate our test instances based on the well known PSPLIB library (Kolisch and Sprecher 1997, url: `http://www.om-db.wi.tum.de/psplib/library.html`). The PSPLIB library contains problem sets for single- and multi-mode resource-constrained project scheduling problems. In particular, we use the PSPLIB single-mode datasets listed in Table 2.2.

No. Projects	**Filename**	**Location:** `www.wiwi.tu-clausthal.de/fileadmin/...`
100	psp1.sh	`...Produktion/Benchmark/RCPSP/testset_ubo100.zip`
200	psp1.sh	`...Produktion/Benchmark/RCPSP/testset_ubo200.zip`
500	PSP1.sh	`...Produktion/Benchmark/RCPSP/testset_ubo500.zip`
1000	PSP1.sh	`...Produktion/Benchmark/RCPSP/testset_ubo1000.zip`

Table 2.2: PSPLIB instances used to construct different instances of the MSVP problem.

Although PSPLIB does not contain instances for the MSVP problem, we use both the precedence and resource constraints provided in its instances. To construct an instance for the MSVP problem, we split the set of projects in a PSPLIB instance into 10 subsets. These subsets represent the time periods in which a project demand resources in the MSVP problem formulation (cf., (2.20)). For example, if a project belongs to the second subset, then this project demands resources in the second time

period in the MSVP problem. This procedure defines the left-hand side coefficients of the resource constraints in (2.20). To vary the complexity of the MSVP instance, we set the right-hand side of the resource constraints to be equal to a fraction of the sum of the left-hand side coefficients. This fraction ranges from the smallest value that results on a feasible instance of the problem to 1.00 (i.e., the resource constraint is redundant).

To generate instances of different sample size, additional samples for the NPVs are generated by adding noise and re-sampling the oil and gas projects' NPVs described in Section 2.4. Following the same procedure as in Section 2.4.2, we also generate different correlations levels among the NPVs. These NPVs correlations range from -1 to 1 as shown in Figure 2.5. Precedence constraints are included without modifications.

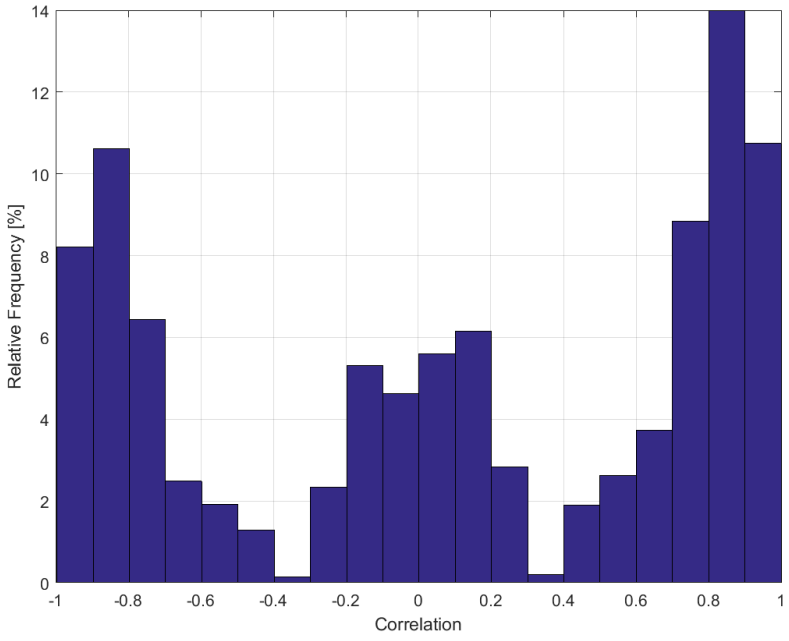


Figure 2.5: Histogram of the NPV correlations (excluding the diagonal) in a general instance of the MSVP problem.

To run our computational experiments, we use MATLAB on a 64-bit workstation with AMD Opteron 2.0 GHz CPU and 32 GB of RAM. We use `CPLEX 12.5` to solve

both the MIQP formulation and the MILP iterations in the Benders-based algorithm. In Algorithm (1), we use $\epsilon = 0.5\%$, and, for fairness of the comparison, we also set the `CPLEX` relative optimality gap to stop the MIQP to $\epsilon = 0.5\%$. We impose a time limit of 3600 seconds for each experiment.

2.5.2 Numerical results

In this section we compare the performance of the Benders-based solution approach described in Section 2.3.2 with the MIQP formulation of the MSVP problem. Table 2.3 shows the results of our experiments for different instances that are generated using the procedure described in Section 2.5.1. In total, we generate 452 instances that include different number of projects and samples, as reported in columns 1–2. For each instance, we use a minimum expected portfolio NPV, μ_0, within the range shown in columns 4–5. Also, the range of the factor that modifies the right-hand-side of the resource constraints is reported in columns 6–7.

To illustrate the variability existing in our test instances, columns 8 and 9 in Table 2.3 summarize the minimum and maximum number of projects selected in the optimal solution of the MSVP problem. In this case we see that variations in the input parameters, besides project and sample size, lead to instances of the MSVP problem with very different solutions. The computational time of the MIQP and the Benders-based solution scheme are reported in columns 10 and 11, respectively. All the tested instances are solved within the time limit, implying that the Benders solution approach obtains the optimal semivariance in the MSVP problem within a 0.5% margin of error. As shown in column 10 (MIQP) and column 11 (Benders), the average solution time of the Benders solution approach is much lower than the MIQP approach. This difference increases as the number of samples in the problem grows. This is not surprising, given that the size of the master problem used in the Benders solution approach does not change with the number of samples. Instead, the number

No.	No.	No.	μ_0		Resource		Projects		Time (s)			
Projects	Samples	Instances	min	max	min	max	min	max	MIQP	Benders	iter.	Speedup
100	1000	32	10	500	0.64	1.00	16	82	1.32	0.15	3.34	8.07
100	5000	20	10	2000	0.70	1.00	11	74	13.08	0.16	2.00	80.07
100	10000	17	10	1500	0.70	1.00	7	79	43.35	0.27	2.00	158.80
200	1000	99	50	10000	0.27	1.00	4	135	3.15	0.13	2.00	23.97
200	5000	29	50	15000	0.70	1.00	4	200	26.42	0.29	2.00	93.32
200	10000	28	50	15000	0.70	1.00	4	200	104.75	0.47	2.00	224.56
500	1000	63	100	10000	0.04	1.00	2	475	12.46	0.93	2.75	15.24
500	5000	19	100	10000	0.70	1.00	21	364	154.31	1.04	2.00	153.73
500	10000	20	100	10000	0.70	1.00	2	457	341.50	1.63	2.32	210.25
1000	1000	86	100	15000	0.04	1.00	3	852	13.89	1.37	2.64	10.87
1000	5000	19	100	15000	0.70	1.00	22	785	263.92	3.28	2.95	87.78
1000	10000	20	100	15000	0.70	1.00	10	746	571.08	8.80	6.20	76.08
Geo. Mean												59.17

Table 2.3: Comparison between MIQP and Benders-based linear solution scheme for general instances of the MSVP problem generated from PSPLIB instances of the resource constrained project scheduling problems. The column Resource indicates the range of the factor used to constraint the resources available in the instance. The column Projects indicate the range of number of projects selected in the optimal solution of the instances. In all instances, differences between the semivariance values of MIQP and Benders algorithms are within a 0.5% margin of error, and the Benders algorithm is faster than the MIQP approach.

of samples only affects the computation of the Benders cuts, which is done through a closed-form calculation. The average number of iterations required by the Benders solution approach and the average solution time speedup are reported in columns 12 and 13, respectively. The reported speedups show that the Benders approach is on average 59 times faster than the MIQP, demonstrating the efficiency of using the Benders solution approach for general large-scale instances of the MSVP problem.

2.6 Concluding Remarks

In this chapter, we studied the MSVP problem. After presenting a convex quadratic formulation of the problem, we proposed two alternative linear solution schemes that effectively solve this problem. These schemes have both practical and computational advantages over a direct MIQP approach to solve the MSVP. The first scheme is based on a MILP approximation that overestimates the projects' portfolio NPV semivariance by solving a single MILP. Aside from providing a formal proof of this overestimation, the computational tests show that the MILP approximation is very accurate when dealing with projects with positively correlated NPVs. Moreover, for instances of the MSVP problem with a moderate number of projects in which it is desired to use a large number of samples to accurately estimate the projects' portfolio NPV semivariance, the MILP approximation solution approach is shown to consistently outperforming the default `CPLEX 12.5` MIQP solver that can be used to directly solve the MSVP problem.

In a more general setting, we proposed a Benders-based linear solution scheme that allows the decision maker to solve the MSVP problem for any positive or negative level of correlation among the NPVs. This approach has proven to be effective, consistently outperforming the default `CPLEX 12.5` MIQP solver for general large-scale instances of the MSVP problem.

The proposed methods have a very broad potential of being applied to other problems. In particular, note that some of the key characteristics of oil and gas project selection problems such as non-divisible assets, skewed NPV distributions, resource and precedence constraints, preference for downside-risk measures, etc., are common to project selection problems in other industries. Also, both linear solution schemes can be easily extended to solve MSVP problems with additional combinatorial constraints which provide real features on the projects

Both linear solution schemes can be easily extended to solve MSVP problems with additional combinatorial constraints which provide real features on the projects (cf. transaction costs (Woodside-Oriakhi et al. 2013), transaction lots (Lejeune 2013), cardinality constraints (Bertsimas and Shioda 2009)). Moreover, recent approaches have focused on efficiently solving the Mean-Variance portfolio allocation problems with integrality constraints (Bertsimas and Shioda 2009, Lejeune 2013, to name a few). Thus, extending the Benders solution scheme to address this type of problems will be a promising topic for future research work.

Chapter 3

Computing near optimal value-at-risk portfolios using integer programming techniques

3.1 Introduction

In the context of portfolio risk and asset liability management, Value-at-Risk (VaR) measures the exposure of a portfolio to high losses. VaR is prominent in current regulatory frameworks for banks (see, e.g., the Basel II and Basel III Accords), as well as for insurance companies (see, e.g., the Solvency II Directive). Thus, VaR is an important and popular tool for risk management in the modern financial and risk management literature (see, e.g., Jorion 2001, Wozabal 2012). Accordingly, the development of risk management methods based on VaR has been the focus of extensive research work (see, e.g., Alexander et al. 2006, Bazak and Shapiro 2001, Darbha 2001, Gaivoronski and Pflug 2004, Kaplanski and Koll 2002, El Ghaoui et al. 2003, Glasserman et al. 2000, Wozabal et al. 2010, Benati and Rizzi 2007, Gneiting 2011).

Although VaR is widely used to measure the risk of a given portfolio of assets, it is

not commonly used as a risk measure in the context of computing optimal VaR portfolios; that is, an optimal risk-reward portfolio allocation using VaR as the risk measure. Instead, other risk measures such as the portfolio return's Variance (cf., Markowitz 1952), and the portfolio loss' Conditional Value-at-Risk (CVaR) (cf., Rockafellar and Uryasev 2000) are more commonly used. This is because, in contrast with the above mentioned risk measures, VaR is non-convex and of combinatorial nature (cf., Gaivoronski and Pflug 2004). As a result, the VaR portfolio problem is inherently difficult to solve (see, e.g., Natarajan et al. 2009).

VaR does not (in general) satisfies the commonly accepted axioms of *coherent* risk measures (cf., Artzner et al. 1999, Rockafellar et al. 2004). On the other hand, VaR satisfies the so-called *natural risk statistic* axioms (Heyde et al. 2006). More importantly, it has been recently shown in Gneiting (2011) that VaR is an *elicitable* risk measure (cf., Bellini and Bignozzi 2013). Loosely speaking, elicitability is related to how accurately a risk measure can be forecasted. More precisely, as discussed in Bellini and Bignozzi (2013), it has been shown that while CVaR is generally considered a better risk measure from a mathematical point of view, it requires a higher number of samples for an accurate estimation (see Daníelsson 2011) and it is less robust than VaR (see Cont et al. 2010).

Because of the computational difficulties of optimally solving general instances of the VaR portfolio problem, different heuristics have been proposed in the literature. In particular, consider the work of Gaivoronski and Pflug (2004), Larsen et al. (2002), Verma and Coleman (2005), Cetinkaya and Thiele (2014). Also, given that the VaR portfolio problem belongs to the general class of *chance constraint* optimization problems (cf., Campi and Calafiore 2005), other approximation approaches that can be used are based on relaxations of the VaR quantile constraint for which probabilistic guarantees can be obtained. In particular, consider the work of Campi and Calafiore (2005), de Farias and Roy (2004), Erdogan and Iyengar (2006).

When the standard *sampling approach* (cf., Rockafellar and Uryasev 2000) is used to model the uncertain asset returns, it is well known (see, e.g., Benati and Rizzi 2007, Feng et al. 2015) that the (resulting) VaR portfolio problem can be solved to optimality by formulating the problem as a mixed-integer linear program (MILP). However, this formulation is difficult to solve with current MILP solvers for instances with medium to large number of assets in the portfolio (see, e.g., Benati and Rizzi 2007). Recently, improvements in the solution of this MILP formulation have been obtained in Feng et al. (2015), by tailoring special *branch-and-cut* techniques to solve the problem, as well as improving the *big-M* values used on its MILP formulation. Although these improvements allow for the solution of VaR portfolio problem instances where thousands of scenarios are used to model the uncertain asset returns, the number of assets that are considered in the portfolio is of the order of 25 assets, similar to Benati and Rizzi (2007). Moreover, their solution approach is useful only when the common total wealth constraint is not considered (Feng et al. 2015, Sec. 5).

We present an algorithm to compute near-optimal VaR portfolios that takes advantage of the VaR portfolio problem MILP formulation and provides a guarantee of the near-optimality of the solution. The algorithm makes a straight-forward use of current state-of-the-art MILP solvers (e.g., `CPLEX` and `Gurobi`). Furthermore, this algorithm can be used to obtain guaranteed near-optimal solutions for instances of the VaR problem with up to a hundred assets and thousands of samples to model the uncertain asset returns. In particular, this allows the use of VaR for strategic asset allocation instead of only tactical asset allocation (e.g., by industry sectors). As a byproduct, we obtain an algorithm to compute tight lower bounds on the VaR portfolio problem that outperforms the algorithms for this purpose proposed by Larsen et al. (2002). These algorithms aim at approximating the optimal solution of the VaR portfolio problem by iteratively constructing appropriate instances of the Conditional Value-at-Risk portfolio problem.

The main contribution of the article in relation to the current VaR portfolio allocation literature is to provide a performance-guaranteed heuristic solution approach for the problem which can be used to address the solution of medium to large-scale instances of the problem. The near-optimal guarantee provided by the proposed algorithm is based on the relation between two alternate formulations of the portfolio problem; that is, between minimum risk portfolios satisfying a reward benchmark and the corresponding maximum reward portfolios satisfying a risk benchmark. It is well-known that these alternate formulations of the portfolio problem are equivalent for the mean-variance portfolio model of Markowitz (1952). Krokhmal et al. (2002, Thm. 3) have shown that this equivalence holds for general convex risk measures and concave reward functions. We also study the relationship between the alternate risk-reward and reward-risk formulations of the portfolio problem for more general risk measures and reward functions. Besides providing the foundation for the proposed algorithm to find near-optimal solutions for the VaR portfolio problem, these results extend the characterization provided by Krokhmal et al. (2002, Thm. 3), and rectify some incorrect statements made in Lin (2009) about alternate formulations of the VaR portfolio problem.

The rest of the article is organized as follows. In Section 3.2, the MILP formulation of the VaR portfolio problem is presented. In Section 3.3, the relationship between the alternate formulations of the portfolio problem is studied for general risk measures and reward functions. These results are used in Section 3.4 to develop the proposed algorithm to find near-optimal solutions for the VaR portfolio problem. In Section 3.5, we illustrate the efficiency of the proposed algorithm by presenting numerical results on instances of the VaR portfolio problem constructed using historical asset returns from the US financial market.

3.2 The MILP formulation of the VaR portfolio problem

The Value-at-Risk (VaR) of a portfolio measures its exposure to high losses. Specifically, for a given $\alpha \in (0,1)$ (typically $0.01 \leq \alpha \leq 0.05$), the VaR of a portfolio is defined as the $1-\alpha$ quantile of the portfolio's losses (cf., Rockafellar and Uryasev 2000); or equivalently as the α quantile of the portfolio's returns. Here, we follow the latter definition (as in Gaivoronski and Pflug 2004).

We begin by formally stating the VaR portfolio (allocation) problem; which aims at minimizing the exposure of the portfolio to high losses while maintaining a minimum expected level of profit. Consider n risky assets that can be chosen by an investor in the financial market. Let $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)^{\mathrm{T}}$ be a random variable in $\mathbb{R}^n$ representing the uncertain returns of the n risky assets from the current time $t = 0$ to a fixed future time $t = T$. Let $x = (x_1, \ldots, x_n)^{\mathrm{T}} \in \mathbb{R}^n_+$ denote a portfolio on these assets; that is, the percentage of the available funds to be allocated in each of the n risky assets. Following Gaivoronski and Pflug (2004), given $\alpha \in (0,1)$, the α-level VaR of the portfolio is defined as follows:

$$\mathrm{VaR}_\alpha(x^{\mathrm{T}}\boldsymbol{\xi}) = \mathbb{Q}_\alpha(x^{\mathrm{T}}\boldsymbol{\xi}), \tag{3.1}$$

where $\mathbb{Q}_\alpha(x^{\mathrm{T}}\boldsymbol{\xi})$ is the α quantile of the portfolio's return distribution; that is, $\mathbb{Q}_\alpha(x^{\mathrm{T}}\boldsymbol{\xi}) = \inf\{v : \Pr(x^{\mathrm{T}}\boldsymbol{\xi} \leq v) > \alpha\}$. Also, the expected portfolio return from $t = 0$ to $t = T$ is given by $\mathbb{E}(x^{\mathrm{T}}\boldsymbol{\xi})$. Above, $\Pr(\cdot)$ and $\mathbb{E}(\cdot)$ respectively indicate probability and expectation.

A (single-period) VaR portfolio problem aims at finding the portfolio $x \in \mathbb{R}^n_+$ to be constructed at $t = 0$, in order to minimize the portfolio's VaR_α, subject to the portfolio having a given minimum expected return μ_0, and possibly satisfying some

linear diversification constraints. Formally, the VaR portfolio problem is:

$$\begin{aligned} \min \quad & -\mathrm{VaR}_\alpha(x^{\mathrm{T}}\boldsymbol{\xi}) \\ \text{s.t.} \quad & \mathbb{E}(x^{\mathrm{T}}\boldsymbol{\xi}) \geq \mu_0 \\ & x^{\mathrm{T}}\mathbb{1} = 1 \\ & x \in \mathcal{X} \subseteq \mathbb{R}^n_+, \end{aligned} \tag{3.2}$$

where $\mathbb{1} \in \mathbb{R}^n$ is the vector of all-ones, $\mu_0 \in \mathbb{R}$ is the given target minimum expected portfolio return, and $\mathcal{X} \subseteq \mathbb{R}^n$ is a given set defined by linear constraints; which are typically used to enforce certain diversification constraints on the portfolio $x \in \mathbb{R}^n_+$. For the moment, it is assumed that no short positions are allowed in the portfolio; which is the most common situation in practice (cf., Michaud 1998).

As discussed in Gaivoronski and Pflug (2004), there are two main approaches to solve (3.2): the *parametric* approach, in which it is assumed that the asset returns are governed by a known distribution ((see, e.g., Lobo 2000), where asset returns are assumed to be normally distributed); and the *sampling* approach, which uses a finite number of samples $\xi^1, \ldots, \xi^m \in \mathbb{R}^n$ of the asset returns (see, e.g., Gaivoronski and Pflug 2004), that are typically obtained from historical data, simulations, or a combination of both. The latter approach is used in the well-known Conditional Value-at-Risk (CVaR) portfolio model (cf., Rockafellar and Uryasev 2000). Here, we adopt the sampling approach, which following Gaivoronski and Pflug (2004, Section

2.1) leads to the VaR portfolio problem (3.2) being written as:

$$\begin{aligned} z_{\mathrm{VaR}} := \quad & \min \quad -\nu \\ & \text{s.t.} \quad \nu = \min^{\lfloor \alpha m \rfloor + 1}\{x^{\mathrm{T}}\xi^1, \ldots, x^{\mathrm{T}}\xi^m\} \\ & \qquad x^{\mathrm{T}}\mu \geq \mu_0 \\ & \qquad x^{\mathrm{T}}\mathbb{1} = 1 \\ & \qquad x \in \mathcal{X} \subseteq \mathbb{R}^n_+, \nu \in \mathbb{R}, \end{aligned} \tag{3.3}$$

where ν represents the $\mathrm{VaR}_\alpha(x^{\mathrm{T}}\boldsymbol{\xi})$, the vector of mean return estimates is, for simplicity, considered to be given by $\mu := (1/m)\sum_{j=1}^m \xi^j$. However, our results are independent of this choice, and a variety of other estimation methods can be used (see, e.g., Black and Litterman 2001, Meucci 2007). Also, for $k \in \{1, \ldots, m\}$, and $u^j \in \mathbb{R}$, $j = 1, \ldots, m$, the k-th smallest element in $\{u^1, \ldots, u^m\}$ is denoted by $\min^k\{u^1, \ldots, u^m\}$ (i.e., $\min^k\{u^1, \ldots, u^m\}$ is the k-th order statistic $u^{(k)}$ in $\{u^1, \ldots, u^m\}$).

Problem (3.3) is equivalent (see, e.g., Benati and Rizzi 2007, Feng et al. 2015) to the following mixed-integer linear program (MILP):

$$\begin{aligned} z_{\mathrm{VaR}} = \quad & \max \quad \nu \\ & \text{s.t.} \quad \sum_{j=1}^m y_j = \lfloor \alpha m \rfloor \\ & \qquad M y_j \geq \nu - x^{\mathrm{T}}\xi^j, \qquad j = 1, \ldots, m \\ & \qquad x^{\mathrm{T}}\mu \geq \mu_0 \\ & \qquad x^{\mathrm{T}}\mathbb{1} = 1 \\ & \qquad x \in \mathcal{X} \subseteq \mathbb{R}^n_+, \nu \in \mathbb{R} \\ & \qquad y_j \in \{0, 1\}, \qquad\qquad j = 1, \ldots, m, \end{aligned} \tag{3.4}$$

where M is a big enough constant (i.e., $M > 2\max\{|\xi_i^j| : i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\}\}$), and as in (3.3), ν represents the VaR of the portfolio $x \in \mathbb{R}^n_+$. The

extra binary variable y_j denotes whether ν is to the right ($y_j = 1$) or to the left ($y_j = 0$) of the sample portfolio return $x^{\mathrm{T}}\xi^j$, for $j = 1, \ldots, m$.

In the literature, it is common to consider two alternate formulations of the portfolio allocation problem. That is, besides the portfolio allocation formulation in which one seeks to obtain the minimum risk portfolio satisfying a reward benchmark (as in Eq. (3.2) above), the alternate formulation in which one seeks to obtain the maximum reward portfolio satisfying a risk benchmark is commonly considered. It is well-known that these alternate formulations of the portfolio problem are equivalent for the classical mean-variance portfolio model of Markowitz (1952) (see, e.g., Krokhmal et al. 2002). Due to the non-convexity of the VaR risk measure, it is not surprising that this equivalence does not hold in general for the VaR portfolio problem considered here. However, the relationship between these two alternate formulations of the VaR portfolio problem is fundamental to develop the algorithm presented here to address the solution of this problem. Below, we formally present the alternate maximum reward portfolio satisfying a minimum VaR risk benchmark $\tilde{\nu} \in \mathbb{R}$.

$$\begin{array}{ll} \max & \mathbb{E}(x^{\mathrm{T}}\boldsymbol{\xi}) \\ \text{s.t.} & -\operatorname{VaR}_\alpha(x^{\mathrm{T}}\boldsymbol{\xi}) \leq \tilde{\nu} \\ & x^{\mathrm{T}}\mathbb{1} = 1 \\ & x \in \mathcal{X} \subseteq \mathbb{R}^n_+. \end{array} \quad . \tag{3.5}$$

Using the sampling approach, and similar to (3.2), problem (3.5) can be reformulated

as the following MILP:

$$
\begin{aligned}
z^*_{\text{VaR}} = \quad & \max \quad x^{\mathrm{T}}\mu \\
& \text{s.t.} \quad \sum_{j \in I} y_j \leq \lfloor \alpha m \rfloor \\
& \qquad M y_j \geq \tilde{\nu} - x^{\mathrm{T}}\xi^j, \qquad j = 1, \ldots, m \\
& \qquad x^{\mathrm{T}}\mathbb{1} = 1 \\
& \qquad x \in \mathcal{X} \subseteq \mathbb{R}^n_+, v \in \mathbb{R} \\
& \qquad y_j \in \{0, 1\}, \qquad\qquad j = 1, \ldots, m.
\end{aligned}
\tag{3.6}
$$

The relationship between the two alternative formulations of the VaR portfolio problems (3.4) and (3.6) will be analyzed in the next section. Moreover, in Section 3.4, this relationship is exploited to obtain approximate solutions of the VaR portfolio problem (3.4) with a near-optimality guarantee.

3.3 On alternate portfolio allocation problem formulations

In portfolio allocation problems one seeks to find the portfolio with minimum risk subject to a constraint on the minimum level of the portfolio's reward. Alternatively, the portfolio allocation problem is also formulated as the problem of obtaining the portfolio with maximum reward subject to a constraint on the maximum level of the portfolio's risk. Similar to Krokhmal et al. (2002), these two problems can be formally and respectively stated as follows:

$$
\begin{aligned}
\beta(a) = \quad & \min \quad \phi(x) \\
& \text{s.t.} \quad R(x) \geq a \\
& \qquad x \in \mathcal{X},
\end{aligned}
\tag{3.7}
$$

$$\begin{aligned} \alpha(b) = \quad & \max \quad R(x) \\ & \text{s.t.} \quad \phi(x) \leq b \\ & \qquad x \in \mathcal{X}, \end{aligned} \tag{3.8}$$

where $x \in \mathbb{R}^n$ represents the portfolio of assets, $\phi(x) : \mathbb{R}^n \to \mathbb{R}$ measures the portfolio's risk, $R(x) : \mathbb{R}^n \to \mathbb{R}$ measures the portfolio's reward, and $\mathcal{X} \in \mathbb{R}^n$ represents the set of admissible portfolios (e.g., $\mathcal{X}$ could contain long only positions constraints or benchmark constraints). Also, $a, b, \in \mathbb{R}$, respectively represent the minimum required reward, and the maximum allowed risk. Throughout, we assume that the set $\mathcal{X} \in \mathbb{R}^n$ is compact (any position on an asset is typically constrained to be within certain lower and upper bounds), and use the usual convention $\beta(a) = +\infty$ (resp. $\alpha(b) = -\infty$) if problem (3.7) (resp. (3.8)) is infeasible.

For the classical mean-variance portfolio allocation model introduced by Markowitz (1952), it is well known that there is a one-to-one correspondence between the optimal portfolios obtained from these two models (i.e., (3.7) and (3.8)). In more generality, it has been shown by Krokhmal et al. (2002, Thm. 3) that this type of one-to-one relationship will hold in more generality whenever the risk measure $\phi(x)$ is convex and the reward measure $R(x)$ is concave.

Not surprisingly, when the risk measure $\phi(x)$ is defined by the portfolio's VaR, there is not a one-to-one correspondence between the portfolio allocation models (3.7) and (3.8). However, as it will be illustrated therein, when using VaR as a risk measure, relaxations of both these problems are useful in addressing the solution of (3.7). Given this, and the fact that Lin (2009) misleadingly shows an example in which there is equivalence between the portfolio allocation problems (3.4) and (3.6), it is worth to study below the relationship between these two problems in a general setting when the risk measure $\phi(x)$ is not necessarily convex and/or the measure of reward $\mathbb{R}(x)$ is not necessarily concave; extending Krokhmal et al. (2002, Thm. 3) to provide

both sufficient and necessary conditions for both (3.7) and (3.8) to have a one-to-one correspondence. These results are formally stated in the remainder of this section.

We define (recall that by assumption $\mathcal{X} \subseteq \mathbb{R}^n$ is compact) the minimum risk and maximum reward that any portfolio in the admissible set $\mathcal{X} \subseteq \mathbb{R}^n$ can have as:

$$\underline{b} = \begin{array}{ll} \min & \phi(x) \\ \text{s.t.} & x \in \mathcal{X} \end{array} \quad \text{and} \quad \bar{a} = \begin{array}{ll} \max & R(x) \\ \text{s.t.} & x \in \mathcal{X}. \end{array} \tag{3.9}$$

Theorem 3.3.1 below, provides sufficient and necessary conditions for a one-to-one correspondence between the portfolio allocation problems (3.7) and (3.8).

Theorem 3.3.1 *Let $I \subseteq [\alpha(\underline{b}), \bar{a}]$ be an interval. The relation $a = \alpha(\beta(a))$ holds for any $a \in I$ if and only if $\beta(a)$ is strictly increasing for all $a \in I$.*

Proof. First notice that $\beta(a)$ is non-decreasing as a function of a. Now, assume that there exists $a_1 \in I$ such that $a_1 < \alpha(\beta(a_1)) =: a_2$. Then $\beta(a_1) \leq \beta(a_2)$ as $\beta(\cdot)$ is non-decreasing, and $\beta(a_2) = \beta(\alpha(\beta(a_1))) \leq \beta(a_1)$ as $\beta(\alpha(b)) \leq b$ for all b. Therefore, $\beta(a_1) = \beta(a_2)$. To prove the other direction, assume $\beta(a)$ is not strictly increasing in I. Then, there exist $a_1, a_2 \in I$ with $a_1 < a_2$ such that $\beta(a_1) = \beta(a_2)$. Then, using that $\alpha(\beta(a)) \geq a$ for all a, we obtain $\alpha(\beta(a_1)) = \alpha(\beta(a_2)) \geq a_2 > a_1$. (See Figure 3.1 for an illustration of the proof.)

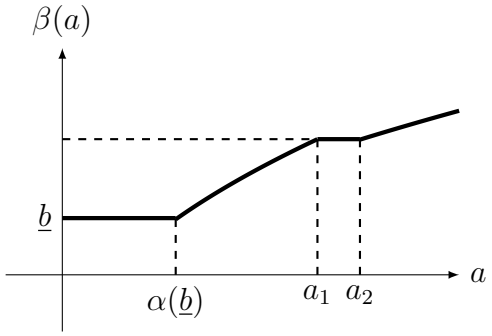


Figure 3.1: Illustration of Theorem 3.3.1.

As mentioned before, it is shown in (Krokhmal et al. 2002, Thm. 3) that convexity in the risk measure, and concavity in the reward, provides sufficient conditions for Theorem 3.3.1 to hold. This result can be obtained as a corollary of Theorem 3.3.1.

Corollary 3.3.2 *Let* $\phi(x) : \mathbb{R}^n \to \mathbb{R}$ *be convex and* $R(x) : \mathbb{R}^n \to \mathbb{R}$ *be concave. Assume* $\{a \in [\alpha(\underline{b}), \bar{a}] : \beta(a) > \underline{b}\}$ *is non-empty and let* $\underline{a} = \inf\{a \in [\alpha(\underline{b}), \bar{a}] : \beta(a) > \underline{b}\}$ *Then* $a = \alpha(\beta(a))$ *for any* $a \in [\underline{a}, \bar{a}]$.

Proof. From Theorem 3.3.1 is enough to show that $\beta(\cdot)$ is strictly increasing on $(\underline{a}, \bar{a}]$. For sake of contradiction, let $\underline{a} < a_1 < a_2 \leq \bar{a}$ be such that $\beta(a_1) = \beta(a_2)$. Let $x_i := \text{argmin}\{\phi(x) : R(x) \geq a_i, x \in \mathcal{X}\}$ for $i = 1, 2$. Thus $\phi(x_1) = \phi(x_2)$. Let $\hat{x}$ be the optimal min-risk portfolio, i.e. $\phi(\hat{x}) = \underline{b}$ and $R(\hat{x}) = \alpha(\underline{b})$. Let $\epsilon := \frac{a_2 - a_1}{2a_2 - a_1 - \alpha(\underline{b})}$. Then $0 < \epsilon < 1$. Let $x' = \epsilon\hat{x} + \epsilon x_1 + (1 - 2\epsilon)x_2$. From the convexity of ϕ, we get that

$$\phi(x') = \phi(\epsilon\hat{x} + \epsilon x_1 + (1-2\epsilon)x_2) \leq \epsilon\phi(\hat{x}) + \epsilon\phi(x_1) + (1-2\epsilon)\phi(x_2) = \epsilon\underline{b} + (1-\epsilon)\phi(x_1) < \phi(x_1).$$

Also, by the concavity of $R(x)$, we get that

$$R(x') \geq \epsilon R(\hat{x}) + \epsilon R(x_1) + (1 - 2\epsilon)R(x_2) \geq \epsilon\alpha(\underline{b}) + \epsilon a_1 + (1 - 2\epsilon)a_2 = a_1.$$

Thus, $x_1 \neq \text{argmin}\{\phi(x) : R(x) \geq a_1, x \in \mathcal{X}\}$, a contradiction.

Krokhmal et al. (2002, Thm. 3) assume a regularity condition for each value of the pair (a, b). In contrast, in Corollary 3.3.2, the ranges of a and b for which the one-to-one correspondence between the alternative formulations holds is precisely characterized.

Although sufficient, the convexity condition in Corollary 3.3.2 is not necessary to have the one-to-one correspondence between the portfolio allocation problems (3.7) and (3.8). To illustrate this, we consider the following simple example in which the risk measure $\phi(x)$ is related to the well-known Huber's function (see, e.g., Huber and

Ronchetti 2009) that commonly appears in robust statistics.

Example 3.3.3 *Let $\kappa > 1$ be given. Let the functions $\phi : \mathbb{R} \to \mathbb{R}$ and $R : \mathbb{R} \to \mathbb{R}$ be given by*

$$\phi(x) = \begin{cases} x^2 & \text{if } |x| \leq \kappa \\ x + \kappa(\kappa - 1) & \text{if } |x| \geq \kappa \end{cases},$$

and $R(x) = x$. Also, let the set $\mathcal{X} = [-2\kappa, 2\kappa]$. The function $\phi(x)$ is not convex, as $2\phi(\kappa) = 2\kappa^2 > (\kappa - 1)^2 + \kappa + 1 + \kappa(\kappa - 1) = \phi(k - 1) + \phi(\kappa + 1)$ (see Figure 3.2 (left)). Thus the conditions of Corollary 3.3.2 are not satisfied. However, it is easy to see that the function $\beta(a)$ is strictly increasing in the domain $a \geq \alpha(\underline{b}) = 0$ (see Figure 3.2 (right)). Note that by changing the domain $\mathcal{X} = \mathbb{R}_+$ and $R(x) = x^2$ one has a similar example where $\beta(a)$ is strictly increasing but now neither $\phi(x)$ is convex nor $R(x)$ is concave.

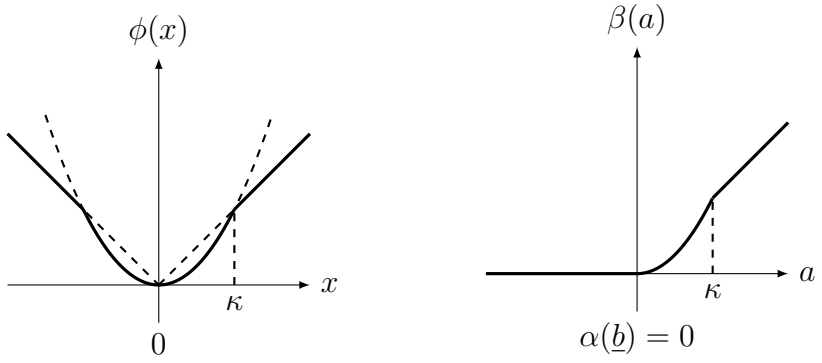


Figure 3.2: Illustration of Example 3.3.3. Function $\phi(x)$ (left), and corresponding $\beta(a)$ (right).

As mentioned earlier, when the risk measure $\phi(x)$ is defined by the portfolio's return VaR, there is in general not a one-to-one correspondence between the portfolio allocation problems (3.7) and (3.8). This is formally stated in the next remark, which corrects the erroneous characterization between problems (3.7) and (3.8) given in Lin (2009).

Remark 3.3.4 *When the risk measure $\phi(x)$ in* (3.7) *is defined by the portfolio's return Value-at-Risk (VaR) $\beta(a)$ is not in general strictly increasing (in the domain $a \geq \alpha(\underline{b})$). This is illustrated with the numerical example below.*

Example 3.3.5 *Consider the instance of Problem* (3.7) *in which $x \in \mathbb{R}^2$ represents the percentage of money invested in the two assets Microsoft* (`MSFT`) *and 3M* (`MMM`). *Let $\mathcal{X} = \{x \in \mathbb{R}^2_+ : x_1 + x_2 = 1\}$. Also, let $\phi(x)$ and $R(x)$ respectively be the estimates of the portfolio's return $VaR_{5\%}$ and expected portfolio return based on a sample monthly returns of* (`MSFT`) *and* (`MMM`), *from April 1986 to December 2006 (source Wharton Research Data Services (WRDS)). After computing $\beta(a)$ in* (3.7) *one obtains Figure 3.3.*

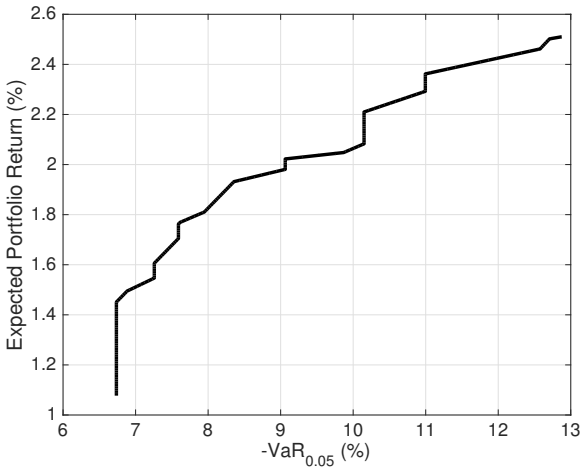


Figure 3.3: Instance of $\beta(a)$ (cf., (3.7)) not being strictly increasing when the portfolio's risk measure is the VaR of the portfolio returns.

Note that the areas of Figure 3.3 in which the risk remains constant while the expected portfolio return increases show that the VaR is not strictly increasing as a function of the expected portfolio return.

We finish this section by showing that one can take advantage of the alternative formulations (3.7) and (3.8) to obtain a measure of the closeness to optimality of a

feasible solution of (3.7) when an appropriate bound on the optimal value of (3.8) can be obtained.

Lemma 3.3.6 *Let $a \leq \bar{a}$. If $\alpha(b) < a$, then $\beta(a) \geq b$.*

Proof. Notice that if $b < \underline{b}$ then by definition $\beta(a) \geq \underline{b} > b$. Thus we assume $b \geq \underline{b}$. For the sake of contradiction assume $\beta(a) < b$. Then there exists $x \in \mathcal{X}$ such that $R(x) \geq a$ and $\phi(x) < b$. Thus x is a feasible solution for (3.8), which implies $\alpha(b) \geq a$, a contradiction.

Proposition 2 *Given $a \leq \bar{a}$. Let $\tilde{x} \in \mathbb{R}^n$ be a feasible solution of (3.7) and $\delta \geq 0$ be a given tolerance. If $\alpha(\phi(\tilde{x}) - \delta) < a$. Then $\phi(\tilde{x}) - \delta \leq \beta(a) \leq \phi(\tilde{x})$.*

In what follows, we use Proposition 2 to provide an algorithm to address the solution of the VaR portfolio problem. As mentioned earlier, in this case, solving the associated minimum risk portfolio problem (3.7) or the maximum return portfolio problem (3.8) to optimality is inherently difficult.

3.4 The algorithm

In this section, we provide an algorithm to obtain approximate solutions for the MILP formulation of the VaR portfolio problem (3.4). First, the goal of the algorithm is to find a near-optimal feasible solution for (3.4) (cf., Section 3.4.1). Next, the goal is to provide a near-optimality guarantee for this feasible solution (cf., Section 3.4.2).

3.4.1 Lower bound for optimal VaR

Let us denote $[m] := \{1, \ldots, m\}$. Now, given $J \subseteq [m]$, let $J^c := [m] \setminus J$, and consider the following problem:

$$
\begin{array}{rll}
\underline{z}_J := & \max & \nu \\
& \text{s.t.} & \displaystyle\sum_{j \in J} y_j = \lfloor \alpha m \rfloor \\
& & M y_j \geq \nu - x^{\mathrm{T}} \xi^j, \qquad j \in J \\
& & 0 \geq \nu - x^{\mathrm{T}} \xi^j, \qquad j \in J^c \\
& & x^{\mathrm{T}} \mu \geq \mu_0 \\
& & x^{\mathrm{T}} \mathbb{1} = 1 \\
& & x \in \mathcal{X} \subseteq \mathbb{R}^n_+, v \in \mathbb{R} \\
& & y_j \in \{0, 1\}, \qquad j \in J.
\end{array}
\tag{3.10}
$$

Note that (3.10) is the optimization problem obtained from (3.4) by setting $y_j = 0$ for all $j \in J^c$. Hence $\underline{z}_J \leq z_{\mathrm{VaR}}$ for all $J \subseteq [m]$. In Algorithm A below, the formulation (3.10), together with an appropriate update of the set J, is used iteratively to obtain near-optimal feasible solutions to (3.4). Specifically, after setting an initial set $J = J_0 \subset [m]$ problem (3.10) is solved. Let $y^J \in \{0, 1\}^{|J|}$ be the optimal value of the binary variables of (3.10). These values are used to construct the linear program below obtained by fixing the binary variables $y \in \{0, 1\}^m$ in (3.4) such that $y_J = y^J$,

and $y_j = 0$, for all $j \in J^c$.

$$
\begin{aligned}
\max \quad & \nu \\
\text{s.t.} \quad & My_j^J \geq \nu - x^{\mathrm{T}}\xi^j, \qquad j = 1, \ldots, m \\
& x^{\mathrm{T}}\mu \geq \mu_0 \\
& x^{\mathrm{T}}\mathbb{1} = 1 \\
& x \in \mathcal{X} \subseteq \mathbb{R}^n_+, \nu \in \mathbb{R}.
\end{aligned}
\tag{3.11}
$$

After solving (3.11), the shadow prices associated with its big-M constraints (the first set of constraints in (3.11)) are used to update the set $J \subseteq [m]$. That is, the set J is augmented by the samples' indices whose corresponding big-M constraints in (3.11) have a positive shadow price. This type of update is similar to the one used when solving MILPs using *branch and price* techniques (see, e.g., Mehrotra and Trick 2007). As described in Algorithm A, this procedure is applied iteratively until no further improvement in the lower bound of (3.4) can be obtained. The VaR of the portfolio obtained at the end of the algorithm serves as a lower bound for the optimal VaR portfolio problem.

Algorithm A Lower bound of optimal VaR.

procedure LOWER BOUND($\alpha, \mu_0, J_0 \subseteq [m], |J_0| \geq \lfloor \alpha m \rfloor$)
  $J \leftarrow J_0$
  $J^{\text{old}} \leftarrow [m]$
  while $J^{\text{old}} \cap J \neq J^{\text{old}}$ **do**
    $y^J \leftarrow \arg_y(\underline{P}_J)$
    $d \leftarrow$ shadow prices of the big-M constraints in (3.10)
    $J^{\text{old}} \leftarrow J$
    $J \leftarrow \{i \in J : y_i^J = 1\} \cup \{i \in J^c : d_i > 0\}$
  end while
  return $\tilde{x} \leftarrow \arg_x$ (3.10) ▷ feasible portfolio for (3.4)
  return $\tilde{\nu} \leftarrow \arg_v$ (3.10) ▷ lower bound for (3.4)
  return $\tilde{y} \leftarrow \arg_y$ (3.10), $I_0 \leftarrow \{i \in [m] : \tilde{y}_i = 1\}$ ▷ to initialize Algorithm B in Section 3.4.2
end procedure

As it will be shown in Section 3.5, Algorithm A provides a tighter lower bound $\tilde{\nu} = \underline{z}_J$, for the VaR portfolio problem (3.4) than those obtained using the CVaR-based algorithms proposed by Larsen et al. (2002) in a comparable running time. More importantly, Algorithm A provides a feasible solution $\tilde{x}, \tilde{\nu}$, for the VaR portfolio problem (3.4) whose near-optimality can be guaranteed using Algorithm B, which is presented in the next section.

3.4.2 Upper bound for optimal return

In this section, the aim is to obtain a measure of the closeness to optimality of the feasible solution $\tilde{x}, \tilde{v}$, for the VaR portfolio problem obtained by Algorithm A. For this purpose, we first apply Proposition 2 to the alternative formulations of the VaR portfolio problem (3.4) and (3.6). Specifically, let $\delta > 0$ be a specified tolerance, and $\tilde{x} \in \mathbb{R}^n_+$ be a feasible portfolio for (3.4), with associated VaR $\tilde{\nu}$; that is, $\tilde{\nu} = \min^{\lfloor \alpha m \rfloor + 1}\{\tilde{x}^{\mathrm{T}}\xi^1, \ldots, \tilde{x}^{\mathrm{T}}\xi^m\}$. Then, from Proposition 2, it follows that if the optimal value of (3.6) satisfies

$$z^*_{\mathrm{VaR}} < \mu_0 \Rightarrow \tilde{\nu} - \delta \leq z_{\mathrm{VaR}} \leq \tilde{\nu}. \tag{3.12}$$

That is, the near-optimality of the feasible portfolio $\tilde{x} \in \mathbb{R}^n_+$ to the optimal portfolio corresponding to the VaR portfolio problem (3.4), can be measured in terms of $\delta \in \mathbb{R}_{++}$.

Clearly, directly solving (3.6) to check whether condition $z^*_{\mathrm{VaR}} < \mu_0$ in (3.12) holds for a feasible portfolio $\tilde{x} \in \mathbb{R}^n_+$ of (3.4) is as computationally inefficient as directly solving (3.4). Therefore, we present an appropriate upper bound for the alternative formulation of the VaR portfolio problem (3.6) that allows to efficiently guarantee the near-optimality of the feasible solutions of the VaR problem obtained after using Algorithm A. Specifically, given $I \subseteq [m]$ with $|I| \geq \lfloor \alpha m \rfloor$ and $\tilde{\nu}$, a lower bound (3.4)

(i.e, $\tilde{\nu} \leq z_{\text{VaR}}$), consider the problem

$$\begin{aligned} \overline{\mu}_I := \quad & \max \quad x^{\mathrm{T}}\mu \\ & \text{s.t.} \quad \sum_{j\in I} y_j \leq \lfloor \alpha m \rfloor \\ & \qquad My_j \geq \tilde{\nu} - x^{\mathrm{T}}\xi^j, \quad j \in I \\ & \qquad x^{\mathrm{T}}\mathbb{1} = 1 \\ & \qquad x \in \mathcal{X} \subseteq \mathbb{R}^n_+, \\ & \qquad y_j \in \{0,1\}, \qquad\quad j \in I. \end{aligned} \tag{3.13}$$

Notice that $\overline{\mu}_{[m]} = z^*_{\text{VaR}}$ (cf., (3.6)). Next, we show that (3.13) is a relaxation of (3.6).

Proposition 3 *Let $I \subseteq [m]$ with $|I| \geq \lfloor \alpha m \rfloor$. Then problem* (3.13) *is a relaxation of* (3.6)*. That is, $\overline{\mu}_I \geq z^*_{\text{VaR}}$.*

Proof. Let $x \in \mathbb{R}^n_+, y \in \{0,1\}^m$ be feasible for (3.6), then we have that $x^{\mathrm{T}}\mathbb{1} = 1$, and $x \in \mathcal{X}$. Moreover, the fact that there exist $y \in \{0,1\}^m$ such that $\sum_{j\in[m]} y_j \leq \lfloor \alpha m \rfloor$, and $My_j \geq \tilde{\nu} - x^{\mathrm{T}}\xi^j$, for all $j \in [m]$, implies that $\tilde{\nu} \leq \min^{\lfloor \alpha m \rfloor + 1}\{x^{\mathrm{T}}\xi^j : j \in [m]\} \leq \min^{\lfloor \alpha m \rfloor + 1}\{x^{\mathrm{T}}\xi^j : j \in I\}$. Thus, $y_I \in \{0,1\}^{|I|}$ satisfies $\sum_{j\in I} y_j \leq \lfloor \alpha m \rfloor$, and $My_j \geq \tilde{\nu} - x^{\mathrm{T}}\xi^j$, for all $j \in I$. Thus, (x, y_I) is a feasible solution for (3.13) with objective value $x^{\mathrm{T}}\mu$.

Notice that from Proposition 3, it follows that

$$\overline{\mu}_I < \mu_0 \Rightarrow z^*_{\text{VaR}} < \mu_0.$$

In Algorithm B below, we take advantage of this fact by iteratively using the formulation (3.13), together with an appropriate update of the set I, with the aim of showing the near-optimality of the feasible solution of the VaR portfolio problem (3.4) obtained from Algorithm A. The set I is updated by heuristically adding samples from the set $[m] \setminus I$ (see, Algorithm B for details) until condition (3.12) is satisfied.

Algorithm B Upper bound for optimal return

procedure UPPER BOUND($\alpha, \beta, \mu_0, \delta, \text{Iter}^{\max}$, and $\tilde{x}, \tilde{\nu}, I_0$ from Algorithm (A))
  $m' \leftarrow \lfloor(\beta\alpha m)\rfloor$
  $I \leftarrow I_0$,
  $\nu \leftarrow \tilde{\nu} + \delta$
  $\overline{\mu}_I \leftarrow \tilde{x}^{\mathrm{T}}\mu$
  while $\overline{\mu}_I \geq \mu_0, I \subset [m],$ and $\text{Iter} \leq \text{Iter}^{\max}$ **do**
    $\overline{\mu}_I \leftarrow$ objective value of (3.13) ▷ $+\infty$ if (3.13) is infeasible
    $x \leftarrow \arg_x (3.13)$ ▷ optimal portfolio for (3.13)
    $I^{\text{old}} \leftarrow I$
    $\nu' \leftarrow \min^{m'+1}\{x^{\mathrm{T}}\xi^j : j \in [m] \setminus I_0\}$
    $I \leftarrow I^{\text{old}} \cup \{j \in [m] \setminus I^{\text{old}} : x^{\mathrm{T}}\xi^j \leq \nu'\}$
  end while
  if $\text{Iter} < \text{Iter}^{\max}$ **then**
    return The δ near-optimality of $\tilde{x}, \tilde{\nu}$ is proven
  else
    return The δ near-optimality of $\tilde{x}, \tilde{\nu}$ could not be verified
  end if
end procedure

3.5 Numerical Results

In this section, we present numerical results to compare the performace of Algorithm A against the CVaR-based algorithms proposed by Larsen et al. (2002) to obtain lower bounds on the VaR portfolio problem (3.4). Moreover, we compare the performance of Algorithm A and Algorithm B to obtain guaranteed near-optimal solutions for the VaR portfolio problem (3.4), against directly solving (3.4) using state-of-the-art mixed integer linear programming (MILP) solvers.

To carry out the experiments we use the daily returns data from Kenneth R. French's website `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french` on 100 portfolios formed on size and book-to-market (10 x 10). The data can be downloaded at `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/ftp/100_Portfolios_10x10_Daily_TXT.zip`. From this data, instances of the VaR portfolio problem (3.4) having number of assets $n \in [30, 90]$, number of samples $m \in [1000, 3500]$ (for every value of n), and expected profit $\mu_0 \in \{\mu^- + \frac{i}{k+1}(\mu^+ - \mu^-) : i = 1, \ldots, k\}$, with $k = 6$

and μ^+ (resp. μ^-) is the largest (lowest) asset mean return. Similar to Benati and Rizzi (2007), Feng et al. (2015), the parameter $\alpha \in (0, 1)$ (cf., (3.1)) is set to the popular value used in practice of $\alpha = 0.01$.

All the code necessary to create the instances of the optimization problems discussed throughout the article is implemented using `Matlab 2016a` and the modelling language `YALMIP`, which is available at `users.isy.liu.se/johanl/yalmip/`. `Gurobi 6.5.0` is used to obtain the solution of all the necessary linear programs and MILPs on a Intel(R) Core (TM) i3-2310M CPU @ 2.10 GHz, 4GB RAM machine.

3.5.1 Lower bound for portfolio's VaR

We compare the performace of Algorithm A against the CVaR-based algorithms proposed by Larsen et al. (2002) to obtain lower bounds on the VaR portfolio problem (3.4).

In all instances, Algorithm A is initialized by setting J_0 as the first $m_0 = \lceil 2\alpha m \rceil$ samples of the instance. Also, the algorithms being compared are allowed to run for a maximum time of up 3600 seconds.

The lower bound results on the VaR portfolio problem (3.4) obtained by the three (3) algorithms are summarized in Table 3.1, Figure 3.4, 3.5, and 3.6. In Table 3.1, for every combination of the number of assets (n) and the number of samples (m), an average is taken over the instances with different values of μ_0, between the values $\mu_0^{\min}$ and $\mu_0^{\max}$. For each algorithm, the column "gap", indicates the relative percentage error between the lower bound on the VaR portfolio problem (3.4) and its optimal solution provided by solving the MILP (3.4). In the few instances when the MILP (3.4) cannot be solved within the maximum allowed time (of 3600 s.), the optimal solution of (3.4) is replaced by the best (higher) lower bound obtained from the lower bound algorithms. Thus, the gap columns in Table 3.1 clearly show that Algorithm A provides tighter lower bounds on the VaR portfolio problem (3.4), than

the ones provided by Algorithm 1 and Algorithm 2 (cf., Larsen et al. 2002). Also, it is clear that the percentage by which Algorithm A provides tighter bounds than Algorithms 1 and 2 is substantial and ranges between $1\% - 7\%$ on average. A more granular evidence of this result is shown in Figures 3.4 and 3.5. In these figures, for each algorithm, the relative gap with respect to the optimal value of the VaR portfolio problem (3.4) for each of the instances considered is plotted in the y-axis, while the x-axis labels indicate the values of the number of samples (m), and expected return (μ_0) of the instance. Also, the number of assets (n) is indicated in each of the plots. In the next section, the tightness of the bounds provided by Algorithm A will be key to be able to guarantee the near-optimality of the solutions for the VaR portfolio problem (3.4) provided by Algorithm A.

As shown in Table 3.1, the tighter bounds obtained by Algorithm A, in comparison with Algorithm 1 and Algorithm 2 in Larsen et al. (2002), are obtained in comparable running times. As mentioned earlier, in Table 3.1, for every combination of the number of assets (n) and the number of samples (m), an average is taken over the instances with different values of μ_0, between the values $\mu_0^{\min}$ and $\mu_0^{\max}$. For each algorithm, the column "T/T^*", indicates the average (over instances with different values of μ_o, and equal n, m) of the ratio between the time taken by each of the algorithms and the minimum of these times on an instance with a particular $\mu_0 \in [\mu_0^{\max}, \mu_0^{\min}]$. From these results it is clear that the average times of the three algorithms are mostly comparable. In Figure 3.6, the average running time information of the algorithms is provided. It is clear from this figure that most of the time, the average time taken by the three algorithms is similar, and that even when there are significant differences between the times, such differences are not of significant practical relevance, since the times required by the algorithms are in the range of at most hundreds of seconds.

Size		μ_0		Alg. A		Alg. 1		Alg. 2	
n	m	min	max	gap	T/T^*	gap	T/T^*	gap	T/T^*
30	1000	0.019	0.058	0.04	1.1	1.23	1.4	0.94	1.2
30	1500	0.043	0.071	0.08	1.4	1.37	1.2	1.48	1.0
30	2000	-0.018	0.056	0.29	1.1	3.02	1.1	2.32	1.0
30	2500	0.012	0.046	0.00	1.1	2.85	1.2	2.56	1.0
50	1000	0.015	0.069	0.00	1.0	2.23	1.2	1.13	1.1
50	1500	0.062	0.075	0.14	1.3	1.59	1.1	1.59	1.0
50	2000	-0.018	0.056	0.29	1.1	2.87	1.1	2.45	1.0
50	2500	0.012	0.068	0.00	1.1	3.04	1.0	3.54	1.0
50	3000	0.005	0.054	0.00	1.4	7.00	1.0	6.22	1.1
50	3500	0.012	0.058	0.00	1.1	1.14	1.0	1.14	1.0
70	1000	0.015	0.069	0.00	1.0	1.50	2.6	2.71	4.3
70	1500	0.039	0.075	0.06	1.3	0.84	1.5	1.30	1.2
70	2000	-0.017	0.061	0.00	1.2	1.96	1.9	1.80	1.8
70	2500	0.012	0.068	0.00	2.1	5.82	2.6	4.78	2.4
70	3000	0.005	0.018	0.00	6.0	4.33	9.3	2.32	1.0
70	3500	0.012	0.069	0.12	2.1	2.94	1.0	2.73	1.1
90	1000	0.007	0.048	0.00	1.1	7.55	1.1	2.37	1.1
90	1500	0.063	0.084	0.00	1.5	0.94	1.0	0.94	1.0
90	2000	0.061	0.061	0.02	1.5	1.68	1.0	0.28	1.0
90	2500	0.012	0.068	0.28	3.4	7.49	1.1	6.28	1.0
90	3000	0.005	0.068	0.00	2.9	2.38	1.5	2.09	1.0
90	3500	0.012	0.069	0.00	3.5	3.12	1.0	3.43	1.1

Table 3.1: Performance of Algorithm A vs. Algorithm 1 and Algorithm 2 in (Larsen et al. 2002) to compute lower bounds on the VaR portfolio allocation problem (3.4). The column gap indicates the VaR lower bound % gap to the optimal VaR. The column T/T^*, is the ratio between the time required to obtain the lower bound T against the minimum time needed by the three algorithms T^*.

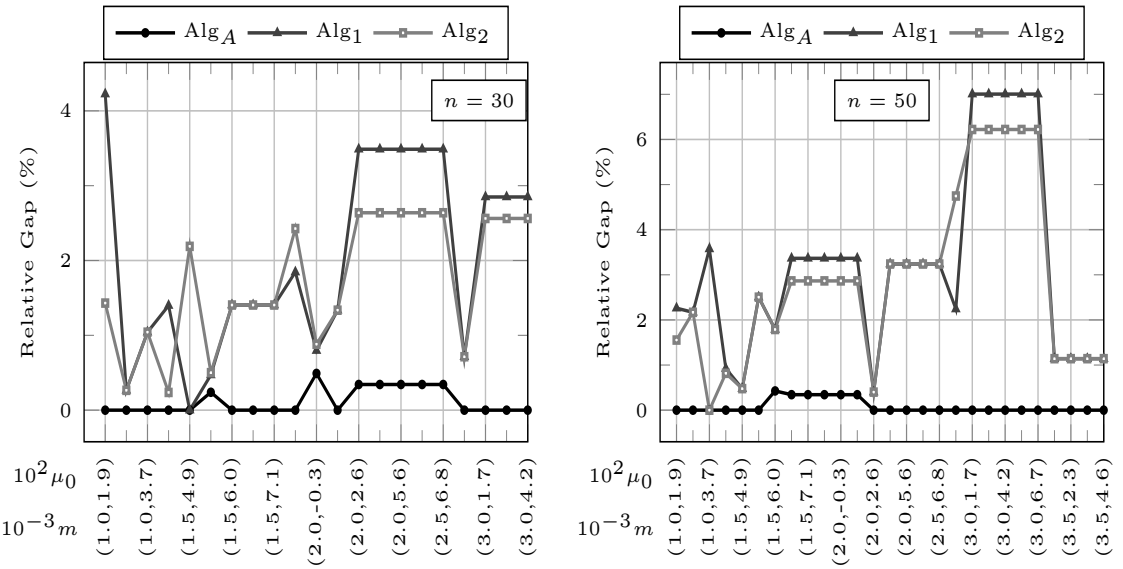


Figure 3.4: Comparison of the relative gap between the optimal value of the VaR portfolio problem (3.4) and the lower bounds for (3.4) provided by Algorithm A, and Algorithms 1 and 2 by Larsen et al. (2002).

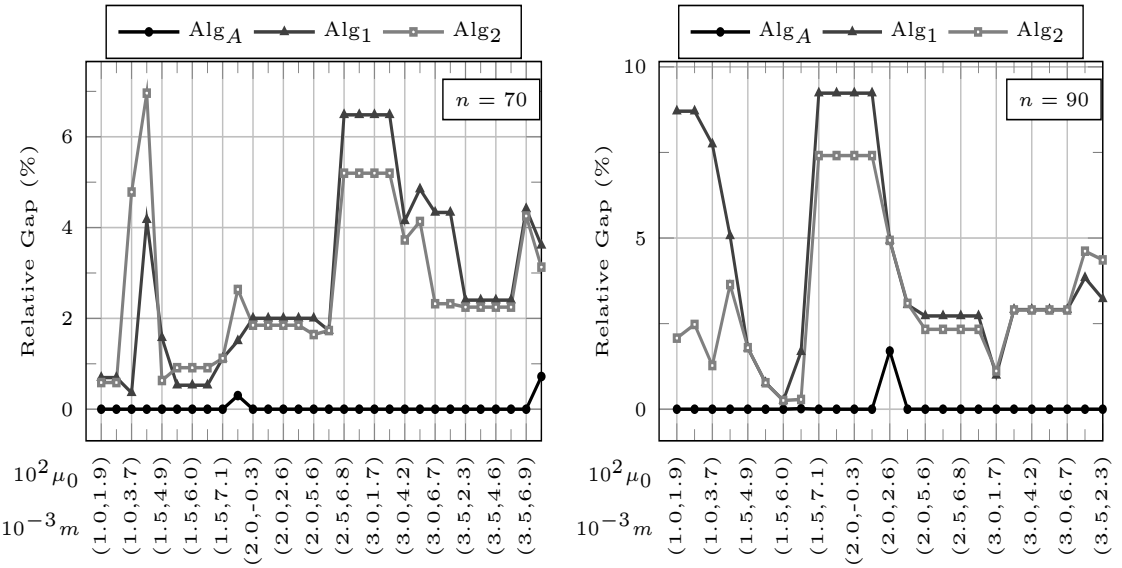


Figure 3.5: Comparison of the relative gap between the optimal value of the VaR portfolio problem (3.4) and the lower bounds for (3.4) provided by Algorithm A, and Algorithms 1 and 2 by Larsen et al. (2002).

3.5.2 Near-optimal VaR portfolio

In this section, we show that by using Algorithm A and Algorithm B, one can efficiently compute guaranteed near-optimal solutions for the VaR portfolio prob-

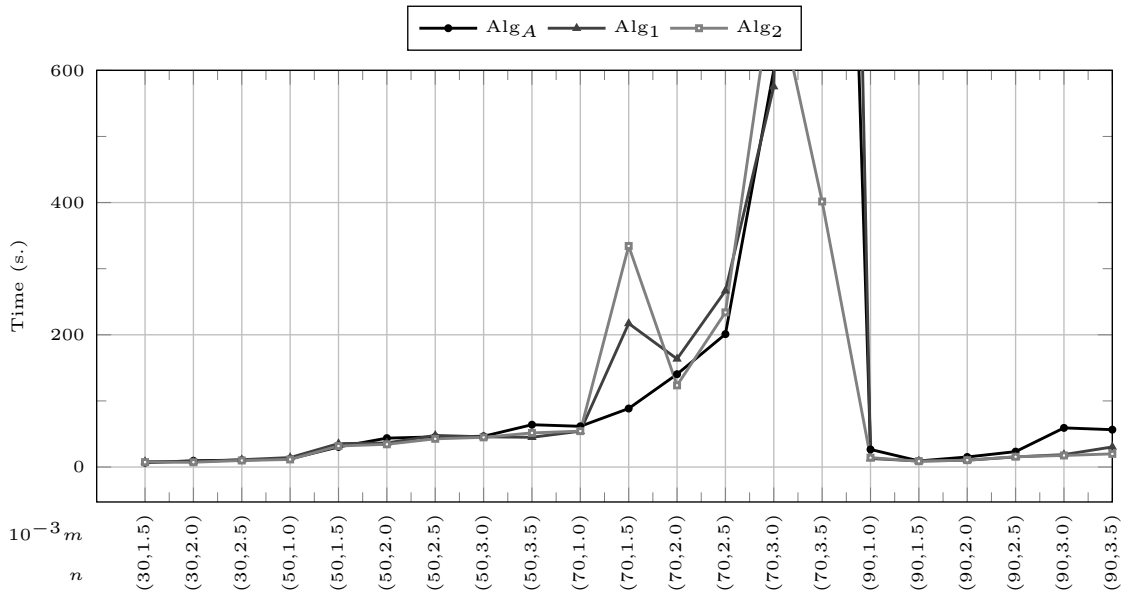


Figure 3.6: Comparison of the average time in seconds needed to run Algorithm A, and Algorithm 1 and 2 in (Larsen et al. 2002) to obtain lower bounds for the VaR portfolio problem (3.4), for instances with given n, m and different values of μ_0 (cf. Table 3.1).

lem (3.4). For that purpose, to obtain the results described below, we first run Algorithm A with J_0 being the first $m_0 = \lceil 2\alpha m \rceil$ samples of the instance. The resulting portfolio $\tilde{x} \in \mathbb{R}^n_+$, VaR lower bound $\tilde{\nu} \in \mathbb{R}$, and the set I_0 (cf., end of Algorithm A) are then used to initialize Algorithm B, Also, we set $\beta = 0.1$, and $\delta = 0.01\tilde{\nu}$. That is, we run Algorithm B seeking to provide a 1% near-optimality guarantee for the portfolio $\tilde{x} \in \mathbb{R}^n_+$. In Table 3.2 and Figure 3.7, the results of finding a near optimal solution to the VaR portfolio problem using Algorithms A and B versus directly solving the MILP formulation (3.4) is compared. For the purpose of brevity, of all the instances considered, Table 3.2 shows, for a particular number of assets n and samples m, the instances in which the MILP solver finds the optimal solution of the VaR problem in the shortest and longest time (depending on the value of μ_0). By comparing the columns "VaR*" and "VaR" in Table 3.2, it is clear that the lower bound on the VaR portfolio problem (3.4) it's equal or very close to the optimal value of the VaR portfolio problem (3.4) (as illustrated also in Figures 3.4 and 3.5). Note

that these lower bounds are well within the 1% desired tolerance. In Table 3.2, T^* is the time taken to solve the MILP formulation (3.4) of the VaR portfolio problem, and T is the time that is taken to obtain a guaranteed near optimal solution for the VaR portfolio problem using Algorithms A and B. Thus, it is clear from the T^*/T columns in Table 3.2 that the latter approach is between 1.2 to 46 times faster than directly solving the MILP formulation. On average, the speed up provided by using Algorithms A and B is approximately of 14 times. Given the time is takes to solve some of the instances of the VaR portfolio, this speed up would be crucial to solve practical instances of the VaR portfolio problem. The effect of the speed up provided by Algorithms A and B can be seen graphically in Figure 3.7, where the time required by Algorithms A and B, versus the time required to solve the MILP formulation of the VaR optimization problem instances in Table 3.2, is shown in a semilogarithmic plot.

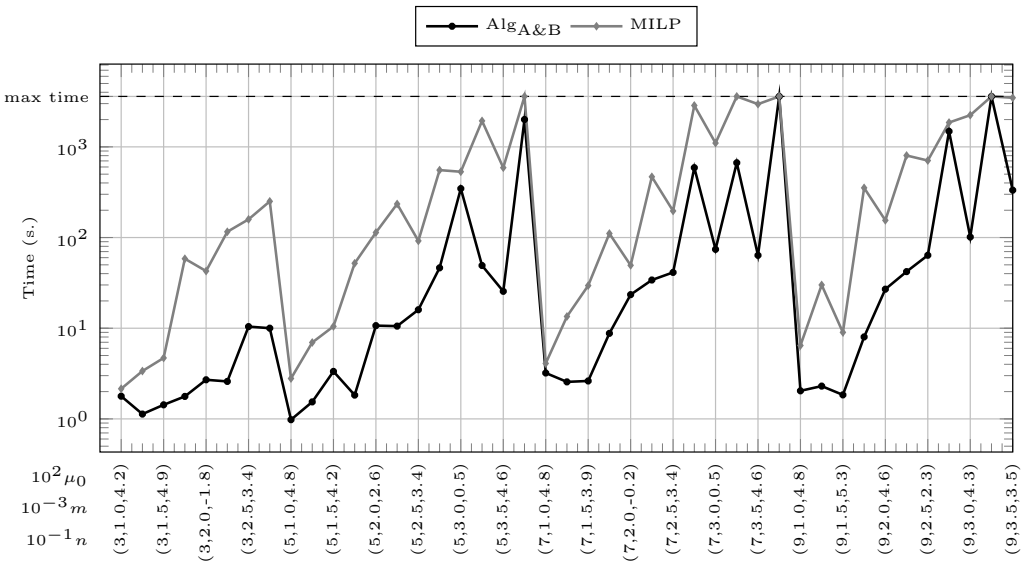


Figure 3.7: Comparison of the time taken by Algorithm (A) and (B) vs directly solving the MILP formulation of the VaR portfolio problem (3.4) for instances with different values of n, m, and μ_0.

			Full MILP		Alg. A & B						Full MILP		Alg. A & B		
n	m	μ_0	VaR*	T*	VaR	T	T*/T	n	m	μ_0	VaR*	T*	VaR	T	T*/T
30	1000	0.058	-2.554	2.2	-2.560	1.78	1.2	70	1500	0.075	-2.002	29.5	-2.008	2.6	11.2
30	1000	0.042	-1.982	3.4	-1.982	1.13	3.0	70	1500	0.039	-1.684	110.8	-1.684	8.8	12.6
30	1500	0.071	-1.968	4.7	-1.968	1.43	3.3	70	2000	0.061	-1.836	49.4	-1.836	23.5	2.1
30	1500	0.049	-1.702	58.4	-1.702	1.77	32.9	70	2000	-0.002	-1.703	467.6	-1.703	34.0	13.7
30	2000	0.026	-1.721	42.7	-1.727	2.7	15.8	70	2500	0.068	-2.121	196.4	-2.121	41.3	4.7
30	2000	-0.018	-1.721	115.8	-1.727	2.59	44.8	70	2500	0.034	-1.876	2860.1	-1.876	589.6	4.8
30	2500	0.057	-1.951	158.7	-1.951	10.42	15.2	70	3000	0.068	-2.024	1101.1	-2.024	74.1	14.8
30	2500	0.034	-1.951	250.6	-1.951	10.01	25.1	70	3000	0.005	***	***	-1.851	667.7	$\star$
50	1000	0.069	-2.449	2.8	-2.449	0.98	2.8	70	3500	0.069	-1.943	2960.4	-1.957	63.7	46.5
50	1000	0.048	-1.973	7.0	-1.973	1.54	4.5	70	3500	0.046	***	***	***	***	***
50	1500	0.075	-2.041	10.5	-2.050	3.34	3.1	90	1000	0.076	-2.325	6.5	-2.339	2.0	3.1
50	1500	0.042	-1.699	52.0	-1.699	1.83	28.5	90	1000	0.048	-1.906	30.0	-1.906	2.3	13.0
50	2000	0.011	-1.721	113.6	-1.727	10.68	10.7	90	1500	0.084	-2.313	9.0	-2.313	1.8	4.8
50	2000	0.026	-1.721	234.5	-1.727	10.55	22.2	90	1500	0.053	-1.669	353.0	-1.669	8.0	44.0
50	2500	0.068	-2.164	91.9	-2.164	15.99	5.8	90	2000	0.061	-1.825	154.8	-1.825	27.0	5.7
50	2500	0.034	-1.920	554.3	-1.920	46.33	12.0	90	2000	0.046	-1.696	803.4	-1.696	42.1	19.1
50	3000	0.067	-2.028	531.7	-2.004	346.59	1.5	90	2500	0.068	-2.115	706.5	-2.115	63.8	11.1
50	3000	0.005	-1.887	1933.9	-1.887	49.15	39.4	90	2500	0.023	-1.826	1855.4	-1.826	1483.7	1.2
50	3500	0.069	-1.982	589.4	-1.984	25.53	23.1	90	3000	0.068	-2.024	2237.0	-2.024	101.4	22.1
50	3500	0.046	***	***	-1.849	2003.1	$\star$	90	3000	0.043	***	***	***	***	***
70	1000	0.069	-2.296	4.1	-2.296	3.21	1.3	90	3500	0.069	-1.931	3479.2	-1.931	333.1	10.4
70	1000	0.048	-1.901	13.5	-1.901	2.56	5.3	90	3500	0.035	***	***	***	***	***
Average Speed Up							14.35	Average Speed Up							13.64

Table 3.2: Comparison of VaR values and running times of full MILP formulation (3.4) vs. Algorithms A & B, for instances of the VaR portfolio problem (3.4) with different no. of assets (n), no. of samples (m), and expected return μ_0. The column T^*/T shows the speed up obtained with Algorithms A & B over solving the full MILP formulation. Instances with "***" where not solved within the 3600s. limit time. The "$\star$" in column T^*/T that ratio cannot be computed.

3.6 Final Remarks

Thus far, we have only considered portfolio allocation problems in which no short positions are allowed (i.e., $\mathcal{X} \subseteq \mathbb{R}^n_+$ in (3.2)). In practice, none of the main characteristics of the MILP formulation (3.4) of the VaR portfolio problem change when considering portfolios were short positions are allowed (i.e., $\mathcal{X} \subseteq \mathbb{R}^n$ in (3.2)). Clearly, only the choice of the Big-M constant M is affected by allowing short positions. However, under the practical assumption that there is $U \in \mathbb{R}_+$ (e.g., due to liquidity) such that $U \geq \max\{i \in \{1, \ldots, n\} : |x_i|\}$, then the M in (3.4) can be set to $M > 2U \max\{|\xi_i^j| : i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\}\}$.

With that said, in this paper, we studied the VaR portfolio selection problem, which is of high relevance in practice, and even in theory, thanks to development of the so-called *natural risk statistic* axioms (Heyde et al. 2006) and the introduction of the concept of *elicitability* (cf., Bellini and Bignozzi 2013) to classify risk measures. To address the inherently difficult task of solving the VaR portfolio problem, here we propose a tandem of approximation algorithms to produce near-optimal solutions to the VaR portfolio problem. This is done by first using Algorithm A to obtain a good feasible solution for the VaR portfolio problem, and as such, provide a lower bound for the optimal VaR associated with (3.4). This algorithm is shown to outperform recent algorithms proposed for this purpose by (Larsen et al. 2002), based on the iterative solution of appropriate CVaR portfolio problems. Then, in Algorithm B, one aims to prove a %1 optimallity guarantee for the feasible solution obtained at the end of Algorithm A. The results obtained here, show that using both Algorithm A and B allows to more efficiently solve VaR portfolio problems with up to a hundred assets and thousands of samples, compared to solving the VaR portfolio problem directly with a MILP solver. This results clearly improve the recent results of (Larsen et al. 2002) on lower bounds for the VaR portfolio problem, and the recent results of (Feng

et al. 2015, Sec. 5) on solving VaR portfolio problems for 25 assets without taking into account the total wealth constraint. Moreover, the proposed algorithms are funded on a study of the alternative formulations of the risk-reward portfolio allocation problem that extends the work done in this area recently by Krokhmal et al. (2002, Thm. 3)

Finally, we believe that the proposed algorithms can be also applied to solve the broader group of chance constrained optimization problems (cf., Sarykalin et al. 2008).

Chapter 4

Systematic prioritization of sensor improvements in an industrial gas supply network

4.1 Introduction

The U.S. industrial gas companies provide indispensable products like oxygen, nitrogen, and hydrogen to manufacturing, health care, transportation, and other essential industries worldwide. In a recent study by the American Chemistry Council, it was shown that industrial gas companies produced approximately $17 billion worth of products in 2010 and employed approximately 60,000 American workers. Furthermore, the study showed that industrial gas companies supply products to industries in the U.S. that account for 42% of America's Gross Domestic Product (Council 2012).

One of the key decision support techniques used by this type of companies is the solution of *industrial gas supply network optimization* models (see, e.g., van den Heever and Grossmann 2003). These models are extremely helpful to identify optimal

operating settings, by expressing real life constraints and conditions mathematically. Also, they are extensively used to describe and integrate all components of the industrial gas supply network within a single framework. However, industrial gas supply network optimization models are often difficult to solve due to the presence of physical and quality constraints, which result in discontinuities and other non-convexities (cf., Nocedal and Wright 2006, Kuhn 2014) on the mathematical formulation of the problem (cf., van den Heever and Grossmann 2003).

There is a vast literature on industrial gas supply network optimization models. For example, consider the work of Almansoori and Shah (2006), Fonseca et al. (2008), Kumar et al. (2010), Ding et al. (2011), Yunqiang et al. (2011), Jiao et al. (2013). Moreover, due to their nature, these models are subject to the presence of uncertainties at various levels. Integrating these uncertainties in industrial gas supply network models has also taken wide attention in the literature (see, e.g. van den Heever and Grossmann 2003, Kim et al. 2008, Almansoori and Shah 2012, Jiao et al. 2012, Lou et al. 2014). However, often times in practice, the uncertainties in the system are disregarded because their integration into the models increases model's complexity which is already very difficult to solve under deterministic assumptions. Alternatively, a common approach to avoid increasing the complexity of solving the industrial gas supply network model is to fix the value of uncertain parameters using a point estimate, in order to obtain a deterministic model that approximates the uncertain model.

While many of the uncertainties that appear in industrial gas supply network optimization models are due to the intrinsically random nature of the input parameters in the network such as flow rate, temperature, and pressure levels, industrial gas supply network optimization models are also affected by uncertainties resulting from incorrect sensor readings in the system. Some of these sensors provide feedback signals which are crucial for the control and efficiency of the network. Thus, erroneous

sensor readings could degrade the performance of the network significantly. Incorrect sensor readings could be due to:

i) Outliers in sensor readings: Sometimes sensors may read a value that happens to be outside of the normal range of operation. This could be caused by an inherent error in the corresponding sensor or a sudden change in the supply network conditions (e.g., a sudden pressure drop). We identify the ones caused by inherent sensor errors, which we name as sensor outliers. Outliers are detected by determining whether an out of range reading is due to sudden changes in the network. This can be verified by other sensors readings (e.g., pressure sensors) in the network. Thus, by looking at the correlation between the out of range readings of associated sensor readings, a system change can be distinguished from an outlier reading due to inherent sensor error in the supply network.

ii) Bias in sensor readings: If the measured signal is shifted by a constant from the actual signal throughout the sensor reading time period, then the sensor has a constant offset or bias. The bias could be negative or positive.

iii) Noise in sensor readings: It refers to the high-frequency error component in the sensor measurements. This type of uncertainty is unavoidable and inherent to every sensor, but it can be improved through maintenance or upgrade.

These incorrect sensor readings can be critical because they are used in measuring the key input parameters of the system. In turn, the aggregate effect of inaccuracies in the model parameters leads to inaccuracies in the optimization model's output. This brings a negative effect on customer satisfaction and puts an unnecessary strain in the industrial gas supply network operation. However, these effects can be mitigated by upgrading and maintaining sensors to improve their reading's accuracy.

In this paper, we focus on the impact of improving sensor reading errors on the model's main output; namely, production costs. From now on, we will regularly

refer to production costs as the output of the model. In particular, we use two key performance indicators to prioritize the improvement of sensors in the network.

i) Key Performance Indicator-1 (KPI-1): It measures the average change in the production costs' value over a time horizon when the sensor reading accuracy is improved. The change in the production costs' value over the time is a result of the presence of outliers and bias in the sensor readings, and it has a direct financial meaning for the company because the elimination of these errors can bring savings on the production cost value.

ii) Key Performance Indicator-2 (KPI-2): This KPI measures the change in production costs' volatility when the sensor reading accuracy is improved. The volatility change in production costs' is due to the amelioration of outliers and random noise in sensor readings.

Production facilities prefer to have low uncertainty in their production systems. Often times these uncertainties cause the total production cost of the system to exceed the defined companies' limited operational budget. Moreover, planning under uncertainty is a difficult task for the companies because plant production levels are very sensitive to input parameters' volatility, as any change in the inputs may cause new production settings in these facilities. Shifting from a defined setting to another could bring extra hidden costs and even infeasibilities in the system. For all these reasons, while the average change in the production cost's value (KPI-1) is crucially important for the company's financial benefits, the change in production costs' variability (KPI-2) is equally crucial.

The faulty sensors affecting production cost accuracy can be addressed by upgrades and maintenance to have more precise readings. However, this is not possible for every single sensor in the system due to limitations on the budget allocated for sensors' upgrade and maintenance. Thus, the purpose of this study is detecting the

sensors in a leading US industrial gas supply network whose inaccuracies have the biggest impact in the supply network.

To detect, identify, and determine the sensor faults, a systematic approach is needed. Traditionally, two ways to deal with sensor faults have been used: preventive maintenance and condition-based maintenance. Preventive maintenance is accomplished by regular checking and calibration of sensors, while condition-based maintenance is based on monitoring a process's real-time condition and automatically detecting sensor faults (Kusiak and Song 2009). Sensor fault detection and identification methodologies have focused on the condition based maintenance aiming at the development of automated sensor fault detection systems, which offer cost advantages over the preventive maintenance systems.

For example consider the work of (Venkatasubramanian et al. 2003, Dunia et al. 1996, Lee et al. 2004, Mehranbod et al. 2005, Guo and Kang 2015). These methodologies construct complex predictive models to replicate actual sensors' behavior. Such predictive models can be constructed based on these methodologies using techniques such as principal component analysis (PCA), neural networks, and bayesian belief networks, among others (see, e.g., Kusiak and Song 2009). However, the resulting models based on these methodologies make it difficult for the user to interpret the underlying relation between the input and output elements.

Here, we follow a similar idea and develop a methodology to approximate the relationship between outputs and inputs in the model by using appropriate sensitivity analysis tools (cf., Cacuci et al. 2005, Saltelli et al. 2008). Sensitivity analysis methods have been frequently applied to chemical process optimization models in the literature (cf., Saltelli et al. 2005, Seferlis and Hrymak 1996). However, to the best of our knowledge, these techniques have not been applied in the context considered here.

After modeling the relationship between model inputs and outputs, we design a

heuristic approach to eliminate sensor malfunctions and approximate the true signal. By integrating the true signal to the constructed predictive model, we calculate the relative change in the system outputs when a sensor is improved by computing the proposed KPIs. We will discuss the methodology in Section 4.3 and in Section 4.4.

The rest of the paper is structured as follows: In Section 4.2, we briefly describe the industrial gas supply network and review some relevant sensitivity analysis methods. The methodology used in this paper is described in Section 4.3. In Section 4.4, the chosen methodology is applied to a simplified gas network model, where the corresponding results can be readily validated. In Section 4.5, we discuss the results obtained after applying this methodology to a real industrial gas network in the US. We conclude the paper in Section 4.6 with some final remarks.

4.2 Industrial Gas Supply Network

The optimization model of an industrial gas supply network can be mathematically formulated as follows (van den Heever and Grossmann 2003)

$$\begin{aligned} \min_{f,p} \quad & y(f) \\ \text{st.} \quad & Af = d, \\ & g_i(f,p) = 0, \quad \forall i \in I \\ & h_j(f,p) \leq b, \quad \forall j \in J \\ & L \leq f \leq U, \\ & p^- \leq p \leq p^+, \end{aligned} \tag{4.1}$$

where A is a matrix representing the structure of the network, f represents the vector of flows in the network and p is the vector of pressures. The objective function minimizes the cost of gas production y as a function of the network flows f (more precisely, flows at production facilities). The first set of constraints $Af = d$

ensures the customer demands and flow balances are satisfied. The second set of constraint $g_i(f,p) = 0, \forall i \in I$ represent physical constraints relating flows and pressures throughout the network. The constraints $h_j(f,p) \leq b, \forall j \in J$ ensure that the model solution satisfies operational quality standards. Finally, in model (4.1), both pressures and flows should remain between allowed bounds.

Model (4.1) is a non-convex, nonlinear deterministic optimization problem, which for real-sized networks is very difficult to solve to optimality, and typically can be only approximately solved. Although we cannot provide here the actual customer and pipeline layout due to confidentiality, Figure 4.1 shows a representative layout of the industrial gas network. The network in the company involves tens of customers and plants, which causes the pipeline flow model to be computationally expensive and takes in the order of minutes to be approximately solved. Referring to Figure 4.1, sensors in the network are typically located at each customer node, each plant node, and at the intermediate nodes where the different pipeline branches intersect. Those sensors read, among others, the gas flow, as well as pressure and CO concentration levels. The industrial gas supply network model is set to work in real time, and it is an important advisory tool for setting the physical plant production levels.

In order to obtain the desired analysis of sensors in the industrial gas network, the first step is to obtain a predictive model that approximates the relationship between the sensor readings used as inputs in model (4.1) (e.g., demands and pressures) and the main output of model (4.1); namely, production costs.

To construct predictive models, several sensitivity analysis methods have been proposed in the literature, which can be divided into two main groups: local sensitivity analysis methods and global sensitivity analysis methods (cf., Borgonovo and Plischke 2015). The local methods provide a measure of the local effect on the model output under small changes of the model inputs (cf., Howard 1988). When the relationship between the inputs and output can be described using a simple differentiable function,

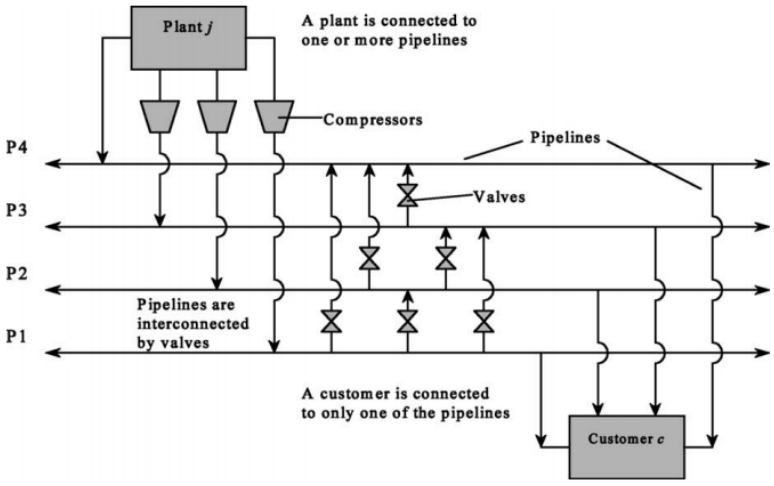


Figure 4.1: Network of plants, pipelines and customers from van den Heever and Grossmann (2003)

.

we can look at the partial derivative of the output function with respect to the input parameters to find out the local impact of the parameters on the model output.

However, the industrial gas supply network model we analyze is non-convex, non-linear, and computationally expensive to solve (e.g., van den Heever and Grossmann 2003). Nevertheless, if one had the access to the exact formulation of the optimization model, sensitivity analysis described in Fiacco (1976) could have been applied to the problem investigated. However, in some circumstances, the exact formulation of the model may not be easily available to the practitioners and the system has to be treated as a black box optimization model where the partial derivatives cannot be directly obtained. Thus, in order to approximate the partial derivatives in a simple and effective way, we use sample historical runs of the optimization model for different parameter settings.

This type of approach based on historical or simulated data to gain information about partial derivatives is typically referred as global sensitivity analysis (cf., Homma and Saltelli 1996, Saltelli et al. 2004).

The total cost of the industrial gas network can be defined as

$$\mathbf{y(x)} = [y^1(x^1), y^2(x^2), \ldots, y^m(x^m))] \tag{4.2}$$

as the functions of uncertain inputs $\mathbf{x} = [x^1, x^2, \ldots, x^m]$, where $x^i = [x_1^i, x_2^i, \ldots, x_k^i]$ represents the values of k parameters in the system. From now on, we denote these uncertain inputs with $\mathbf{x}$. In turn, uncertainty in the $\mathbf{x}$ parameters results in a corresponding uncertainty in the output $\mathbf{y(x)}$.

Different sensitivity analysis techniques will do well on different types of problems. The important aspect here is choosing the most suitable methodology to determine a predictive model between input parameters and output costs. For linear models, linear relationship measures like partial correlation coefficients (PCC) (cf., Brown and Hendrix 2005) will be adequate. For nonlinear but monotonic models, measures based on rank transforms like partial rank regression coefficient (PRCC) (cf., Geladi and Kowalski 1986) will perform well. For non-linear and non-monotonic models, methods based on decomposing the variance of the output is the best choice. One of the examples of these methods is the Sobol's method (cf., Sobol 2001).

Here, we chose Regression Coefficients (RC) (cf., Wagner 1995) as the tool to estimate the desired predictive model. The magnitude of the coefficients in this predictive model will be the indicators for identifying the key input parameters in the system. Favorable results of this least squares approach can be verified in Section 4.4.

After identifying the key parameters in the industrial gasses optimization model using an appropriate sensitivity analysis, we need to do further analysis to measure the impact of the errors in sensor readings. For this purpose, we develop a series of heuristic approximation methodologies to eliminate these sensor reading errors step by step that helps us to conduct the sensor improvement analysis in the network.

4.3 Methodology

A simple approach to address the problem considered here is to upgrade a sensor based on its current reading accuracy which aims to reduce only process variation. Specifically, this approach allocates a budget so that the sensors with the highest inaccuracies are upgraded to decrease model output's variation. However, this would not necessarily be the best way to eliminate errors from the system because two main reasons. First, high input uncertainty does not always lead to high output uncertainty (in our case the production cost). Second, this simple approach also fails to consider the importance of the evaluation criterion KPI-1, by ignoring the complex impact of the sensor reading errors in the total production costs.

Here we propose a methodology which is suitable to work in real time and with scarce data, that provides an effective advisory tool to make decisions regarding the improvement of sensors in the network within the budget limitations. The historical data of the network in the company's database is enormous. However, the older the data gets, the more meaningless it becomes. Thus, the methodology has to be run with recent data.

Here provide a brief explanation of each step of the methodology represented in Figure 4.2.

i) The first step is to generate samples of the input parameters $\mathbf{x} = [x^1, x^2, \ldots, x^k]$ using simulation or by obtaining historical data from previous runs from the industrial gas network model. In our case study problem in Section 4.4, we generate the samples using Monte Carlo simulation while we use historical data from the gas supply network in the real case implementation given in Section 4.5.

   After collecting the data, we run the following two steps in parallel.

ii) A predictive model is constructed to approximate the relationship between production costs and input parameters. The least squares approach provides a good

approximation to get the desired sensor improvement prioritization in the case study problem. However, due to the possible multicollinearity between the input parameters in the real system, we apply the ridge regression (cf., Hoerl and Kennard 1970) to consider this factor. The ridge regression technique basically penalizes the size of regression coefficients due to multicollinearity in the model.

iii) A heuristic methodology for sensor reading error elimination is applied. The heuristic methodology is applied as a univariate analysis, i.e. it is applied to a single sensor at a time. For the selected sensor reading, the heuristic starts by looking at the outliers to eliminate. Note that outliers definitely produce an increase in the production cost's volatility and may cause an increase in the production cost's value depending on the outlier's position, and input-output relationship. After detecting and eliminating the outliers, we eliminate the constant bias in the selected sensor reading. Constant bias in sensors can bring a substantial amount of extra cost in the system. Since it has a financial impact on the output, this needs to be certainly handled in the analysis. Finally, the heuristic approach eliminates the noise in sensor reading. Every sensor noise is assumed to follow a Gaussian distribution $\mathcal{N}(0, \sigma)$ with mean 0 and standard deviation σ. Noise in sensors can cause extra volatility in the system outputs (e.g. production cost, optimal production settings of the plants), and thus it brings up some indirect financial impact to the company.

iv) After running the predictive model and the three step sensor reading error elimination heuristic to approximate the real sensor signal, we reproduce the production costs with the improved signal.

v) In the final step of the model, we calculate two key performance measures previously defined as KPI-1 and KPI-2 that are used to support the decision-making process. These KPIs help us to capture the marginal contribution of a single

sensor reading error, and assign the priorities to the sensors for upgrading purposes. KPIs are calculated based on the comparison of the original production cost values and the regenerated production cost values for every improved sensor one at a time.

To validate the methodology, we first apply it to a case study problem which is introduced in Section 4.4. After it's validation with the simplified network, in Section 4.5, we show how this methodology together with some adaptations can be applied to a real industrial gas network.

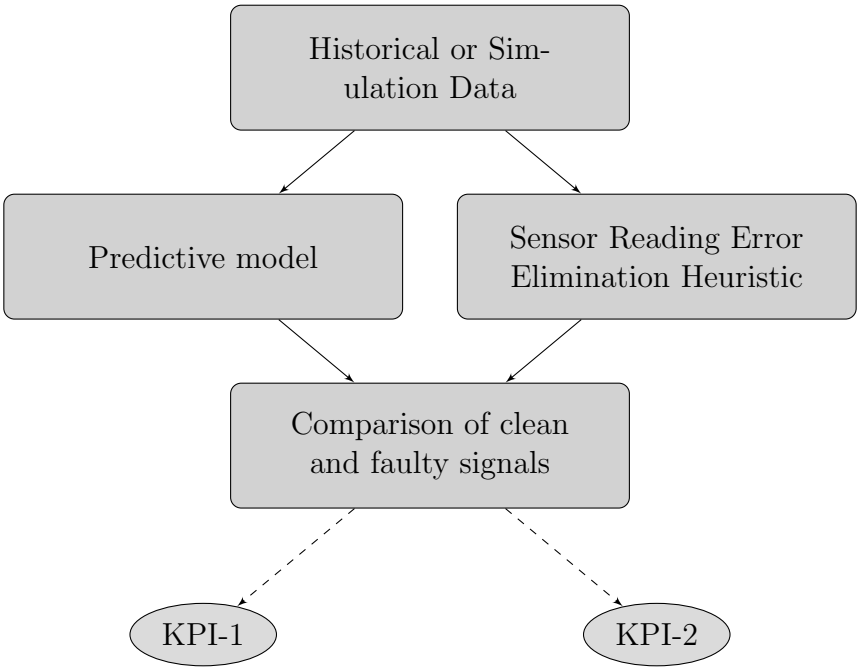


Figure 4.2: Sensor Fault Detection Process Flow Chart

4.4 Case Study Problem

In this section, we illustrate the details and the results of applying the aforementioned sensitivity methodology by implementing it to a simplified version of an industrial gas network model. This allows providing a clear description of the proposed

methodology and validates the approach by providing the desired results regarding the prioritization of sensor improvements in an industrial gas supply network.

4.4.1 Problem Setting

Consider a simplified version of the industrial gas network depicted in Figure 4.3, where there are three plants (P-1, P-2, P-3) and three customers (C-1, C-2, C-3) with the decision variables and parameters described in Tables 4.1 and 4.2 respectively.

f_i	Flow variables in the pipeline	$i = 1 \dots 7$

Table 4.1: Decision Variables

D_j^t	Random demand of customer j at time t	$j = 1 \dots 3$, $t = 1 \dots T$
U_i	Upper bound on f_i	$i = 1 \dots 7$
L_i	Lower bound on f_i	$i = 1 \dots 7$
β_i	Linear objective coefficients for each flow variable	$i = 1 \dots 7$
α_i	Quadratic cost coefficients for each flow variable	$i = 1 \dots 7$

Table 4.2: Input Parameters

In the simplified network, the customers submit their demands to the system which are then fulfilled at the lowest possible cost while satisfying the flow balance and flow bound constraints. The customer demands define the flow rates sent to the customers in the system. In this case, consider the following optimization model.

$$\begin{aligned} \min \quad & \sum_{i}^{7} \alpha_i f_i^2 + \beta_i f_i, \\ \text{st.} \quad & f_1 + f_6 - f_7 = D_1^t, \\ & f_4 + f_2 + f_7 - f_6 - f_5 = D_2^t, \\ & f_3 + f_5 - f_4 = D_3^t, \\ & f_i \in [L_i, U_i], \quad \{i = 1 \dots 7\}. \end{aligned} \tag{4.3}$$

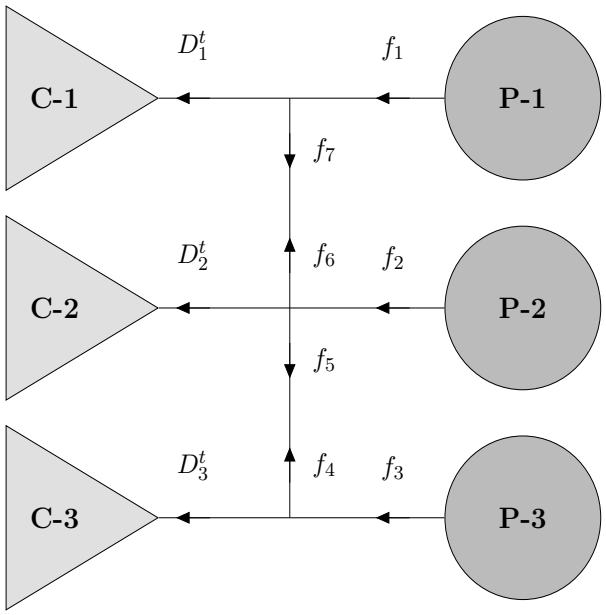


Figure 4.3: Simplified industrial gas network model with three (3) plants (P-1, P-2, P-3) and three (3) customers (C-1, C-2, C-3). Arrows indicate the directions of gas flows.

The first three constraints in formulation (4.3) are flow balance constraints (i.e., $Af = d$ formulation (4.1)). The last constraint set represents the bound constraints on the flows. Formulation (4.3) described above is a simplified version of model (4.1) where constraints defining physical and operational quality standards are disregarded, and there are no pressure constraints. The sensors subject to the analysis are chosen to be the flow rate sensors at customer nodes. Sensors measuring pressure levels at demand nodes are used only as a system check. The objective function is assumed to be convex and quadratic (i.e., $\alpha_i \geq 0$, $i = 1 \dots 7$). This setting allows us to obtain sample data for the global sensitivity analysis easily by solving the model with state of the art optimization solvers.

4.4.2 Problem Data

After setting up the problem, we simulate the data for the input parameters of model (4.3), which is the first step of the flow chart in Figure 4.2.

The values of the input parameters, except for the demand parameters (cf., Section 4.4.2), are defined as follows: $\alpha = [10, 5, 2, 0, 0, 0, 0]$, $\beta = [10, 5, 2, 0, 0, 0, 0]$, $L = [0, 0, 0, 0, 0, 0, 0]$, $U = [150, 70, 50, 10, 10, 10, 10]$ where α is the quadratic cost coefficient vector in \$ per square unit of flow, and β is the linear cost coefficient vector in \$ per unit of flow. These values have the following characteristics that will be the key to illustrate the effectiveness of the proposed approach.

i) The most expensive plant is plant P-1 while the cheapest one is plant P-3.

ii) Similarly, plant P-1 has the largest production capacity while plant P-3 has the least.

iii) Note that the α and β coefficients are non-zero only for flows f_i, $i = 1, 2, 3$ corresponding to plant productions. This means that the cost only increases with the production levels in plants.

iv) There is a limit on the demand of customers that can be supplied from non-adjacent plants. For instance, customer C-1 can only receive a certain amount of gas from plant P-2 and plant P-3. The main supplier of customer C-1 is plant P-1. This also applies to the other customer demands. In particular, flow values f_i, $i = 4, 5, 6, 7$ in the center pipeline are bounded above by 10 units.

Customer demand simulations

In order to simulate the customer demand, we use Auto-Regressive (AR) processes which are commonly used to forecast demands in industrial gas networks. These simulated demand profiles are shown in Figure 4.4. The AR(1) model specified for customer C-1 is given by the following equation

$$D_1^t = 1 + 0.99 D_1^{t-1} + \epsilon_1^t, \tag{4.4}$$

where $\epsilon_1^t \sim \mathcal{N}(0,1)$, $D_1^t = 130.53$, the average value of the customer C-1 demand is given by $\mu_1 = 138.11$ and the standard deviation of the demand is provided as $\sigma_{(D_1^t)} = 3.360$. The AR(1) model specified for customer C-2 is given by the following equation

$$D_2^t = 1 + 0.95D_2^{t-1} + \epsilon_2^t, \tag{4.5}$$

where $\epsilon_2^t \sim \mathcal{N}(0,1)$, $D_2^0 = 60.67$, the average value of the customer C-2 demand is given by $\mu_2 = 62.03$ and the standard deviation of the demand is provided as $\sigma_{(D_2^t)} = 2.360$. The AR(2) model specified for customer C-3 is given by the following equation

$$D_3^t = 0.7 + 0.7D_3^{t-1} + 0.25D_3^{t-2} + \epsilon_3^t, \tag{4.6}$$

where $\epsilon_3^t \sim \mathcal{N}(0,1)$, $D_3^0 = 14.29$, the average value of the customer C-3 demand is given by $\mu_3 = 17.74$ and the standard deviation of the demand is provided as $\sigma_{(D_3^t)} = 2.835$.

The Pearson correlation coefficients of the simulated demand profiles are $\rho_{12} = 0.278$, $\rho_{13} = 0.005$ and $\rho_{23} = 0.137$.

Customer Flow Sensor Reading Simulations

To simulate the errors due to sensor readings of demand profiles, we add some simulated out of range readings which are the candidates to be the outliers, constant bias B and Gaussian noise $N_i^t \sim \mathcal{N}(0, \sigma_i)$ to the demand profiles. That is, we set

$$\tilde{D}_i^t = D_i^t + O_i^t + B_i + N_i^t \quad i = 1, 2, 3, \tag{4.7}$$

where $\tilde{D}_i^t$ represents the sensor readings, O_i^t is the vector of out of range readings suspected to be outliers, $N_i^t \sim \mathcal{N}(0, \sigma_i)$ with $\sigma = [0.5, 1, 1.5]$ are the sensor noises,

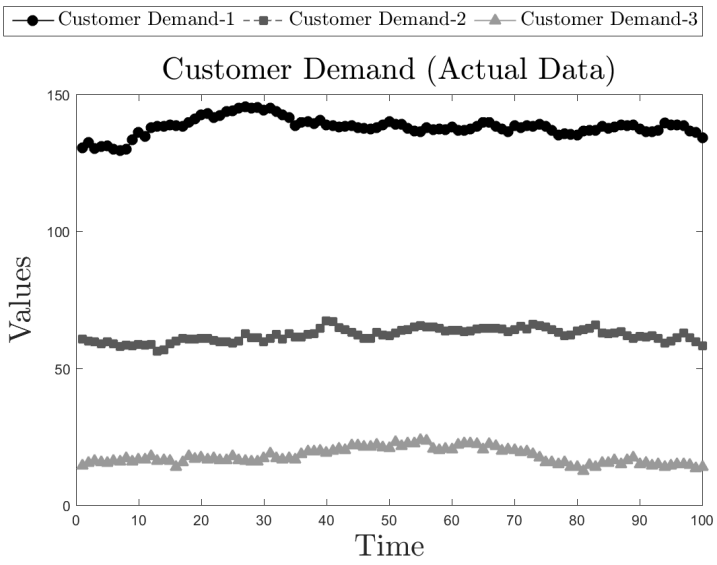


Figure 4.4: Simulated base customer demands.

and $B = \{2, 4, 6\}$ is the constant bias vector.

The resulting sensor reading data can be seen in Figure 4.5, where potential outliers are highlighted by big markers over data points.

As it can be noticed from the parameter values in (4.7) and Figure 4.5, customer C-3 has the greatest sensor reading errors in terms of all three different error types. On the other hand, customer C-1 has the most accurate sensor reading among the other customers. In these conditions, one would expect to prioritize improving the sensor readings for customer C-3 first, then customer C-2, and finally customer C-1. However, this would be deciding based on the input errors only. To consider the impact of input errors on the output, we need further analysis.

Operating the system by these faulty signals observed in Figure 4.5 may cause large deviations and higher values in the production costs. These deviations can be inspected by comparing the plots displaying production costs produced with the sensor reading data given in Figure 4.6a, and the production costs values produced with the actual demand data given in Figure 4.6b. The real life optimization model

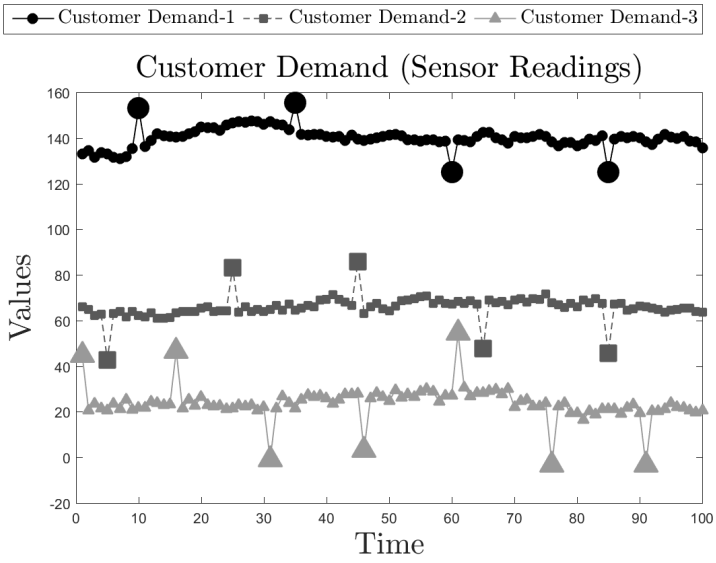
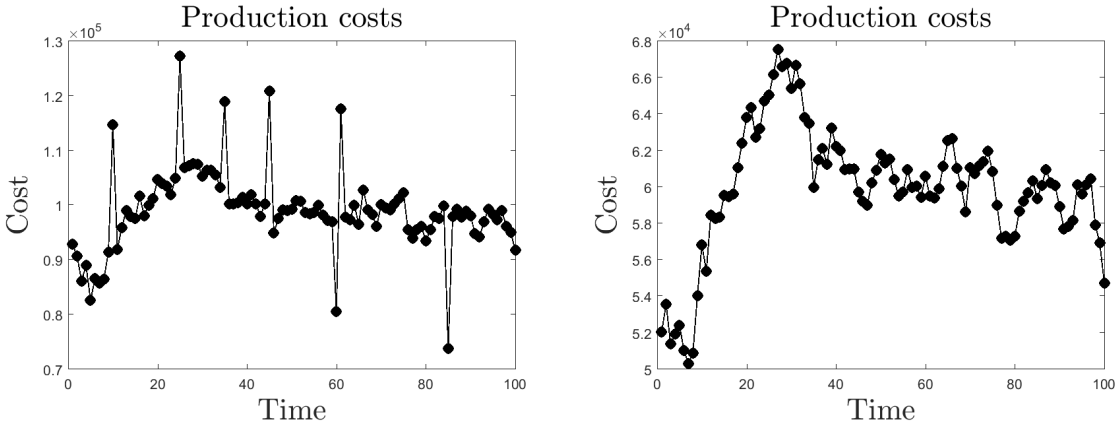


Figure 4.5: Simulated customer demand sensor readings.



(a) with the sensor reading data ($\tilde{D}$). (b) with the actual demand data (D).

Figure 4.6: Cost function output through optimization model.

is computationally very expensive; for this reason, we need a simple way to determine the affect of input errors on the production cost. As mentioned before, this is why sensitivity analysis tools are used to construct a predictive model between model's inputs and outputs. We will compare our findings by controlling simulated sensor reading errors, and show that the proposed methodology estimates the correct order of the sensors for improvement.

4.4.3 Predictive Model

In this section, we apply the second step of the developed methodology in Figure 4.2. After collecting the data in (4.2) for the vector of uncertain inputs denoted by same symbol as in (4.2), and populating the sensor readings of customer demands $\tilde{D}$, we solve model (4.3) T times with uncertain input parameters $\tilde{D}$ to optimality with current optimization solvers (Boyd and Vandenberghe 2004). In particular, we use the `quadprog` solver in a `MATLAB` environment (Coleman et al. 1999). The time required to solve model (4.3) with 3 customers and 3 flow sensors is only of about 0.01 seconds.

The resulting production costs $y(\tilde{D})$ are illustrated in Figure 4.6a, where we set $T = 100$ as the number of observations in the data set. In the real network, this number corresponds to the daily amount of data the company collects from the optimization model. We work with daily data because the system conditions can vary considerably for longer periods. Such high variations cannot be captured in a linear predictive model.

By using the input data created from the simulation of $\tilde{D}$ and the output of the optimization model $y(\tilde{D})$, we get the following linear equation as our predictive model

$$\hat{y} = -67044 + 1208.6 * D_1 + 445.65 * D_2 + 138.10 * D_3 \tag{4.8}$$

where $\hat{y}$ represents the estimated production cost values. Since the multicollinearity is not an issue in the case study problem, ridge regression model coefficients would be expected to be identical to linear regression model coefficients for any ridge parameters.

The residuals of the regression model also satisfy statistical independence, homoscedasticity, and normality assumptions.

According to the regression equation (4.8), the relative importance of the input customer demands are sorted from more to less important as the demand of customer

C-1, the demand of customer C-2, and the demand of customer C-3. This is expected because of the flow restrictions in the pipelines. Customer C-1 has to satisfy most of its demand by using the most expensive gas which is produced in plant P-1 while customer C-3 uses mostly the least expensive gas produced in P-3.

4.4.4 Sensor Reading Error Elimination Heuristic

After defining the predictive model, we run the heuristic approach to eliminate each type of sensor errors step by step. As discussed before, the predictive model and the heuristic for sensor reading error elimination are independent processes as shown in Figure 4.2. Moreover, the heuristic approach is a univariate analysis that is applied to each sensor individually.

Outlier Elimination

The first step of the heuristic approach for sensor reading error elimination is the outlier elimination. The out of range readings are often observed as unusual spikes in flow readings. As mentioned before in Section 4.1, these inaccuracies may happen because of one of the following reasons: due to an inherent error in the sensor reading, or due to an actual sudden change in conditions in the supply network (i.e., like sudden pressure drops). In the first case the error can be reduced by upgrading the sensor. In the latter case, such sudden changes in the network can be determined by other sensors in the network. Thus, by looking at the correlation between the out of range readings of appropriate groups of sensors, outlier readings due to reading errors can be distinguished from out of range readings due to sudden changes in the supply network. Here, we look at the pressure reading values given that pressure and flow rates are highly correlated in industrial gas supply networks.

To detect the potential outliers, we use the methodology called Principal Component Pursuit (PCP) analysis. PCP optimally decomposes a data matrix as the sum of

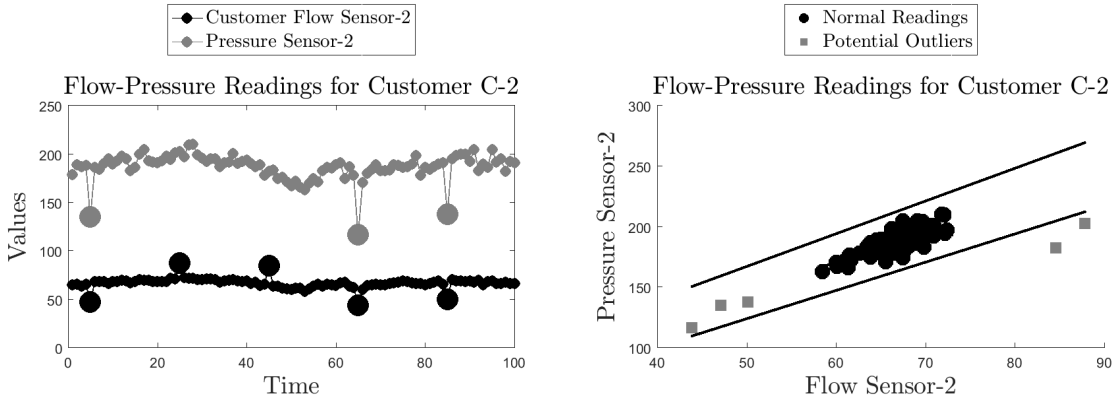
a low-rank matrix and a sparse matrix. Given a data matrix H, PCP is the solution of the convex optimization problem,

$$\begin{aligned} \min \quad & ||Z||_* + \lambda ||E||_1, \\ \text{st.} \quad & H = Z + E, \end{aligned} \tag{4.9}$$

where $||Z||_*$ is the nuclear norm; that is, the sum of the singular values of Z, and $||E||_1$ is equal to the sum of the absolute values of the elements of E. Under certain conditions, and with a Lagrange multiplier λ, the optimization problem (4.9) recovers a low-rank matrix Z corresponding to a fault-free process condition and a sparse matrix E that has non-zero entities corresponding to sensor and process faults, which can be considered as sensor reading abnormalities or sharp changes in the system (cf., Isom and LaBarre 2011). The matrix H is the selected flow sensor reading data vector $\tilde{D}$ in our case. The resulting nonzero values in the E vector helps us to identify out of range readings.

After implementing the PCP routine above, we can identify the out of range readings in the sensor signal similar to those illustrated in Figure 4.5. After identifying the out of range readings, we look at the correlation of these sensor readings with pressure sensor readings associated with the same customer. If the sharp spikes in flow sensor readings are caused by a sudden change in the network, these secondary sensors for pressure readings will also have abnormal readings.

Let's illustrate this process by looking at the correlation between the flow sensor readings and the pressure readings of customer C-2. In Figure 4.7a, two different signals for customer C-2 are shown: the sensor readings for customer C-2 demand, and the simulated pressure sensor reading for customer C-2. Obviously, the marked points in the flow readings are strong candidates for outliers. However, notice that in a few of these points, the pressure level also has abnormal spikes. These spikes occur



(a) Flow-pressure reading plot.

(b) Flow-pressure scatter plot.

Figure 4.7: Simulated flow-pressure plots for customer C-2

at the exactly the same timestamps that the spikes occur in the flow readings. These out of range readings are strong indicators of the sudden changes in the system, and they shall not be classified as sensor reading errors.

This can be better viewed by inspecting the scatter plot between the pressure and the gas flow in Figure 4.7b. The PCP routine identifies out of range readings in the selected flow sensor, and marks them as square dots. Dark circle dots are marked as normal readings for both pressure and flow sensors. We fit a least squares line between these flow and pressure readings excluding the out of range observations. Then, a 95% confidence interval is built around this least squares line. If the out of range readings are beyond these confidence intervals, then they are marked as sensor outliers, and we replace the sensor readings with the values from the low-rank matrix Z for these corresponding points. In Figure 4.7b, the outlier points are the two square dots at the right side of the graph. These points are identified to be sensor reading errors, and needed to be eliminated from the sensor signal. On the other hand, if some of these out of range readings fall between the confidence intervals provided in Figure 4.7b, they are treated as sudden changes in the conditions of the supply network, not as outliers. In Figure 4.7b, the out of range readings caused by system

changes are denoted by the three square dots at the left side of the graph. Notice that they fall between the two confidence interval lines.

The same procedure is also applied to customer C-1's and customer C-3's sensors, and resulting outliers are eliminated from the sensor readings. Figure 4.8 shows the demand profiles after elimination of the outliers caused by inherent sensor errors.

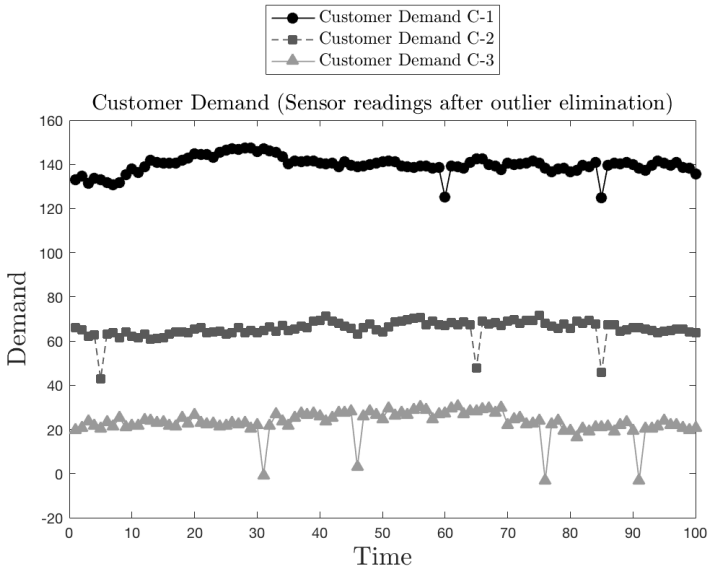


Figure 4.8: Customer demand data (without outliers).

Constant Bias Elimination

Constant bias elimination is the second step of the heuristic methodology to eliminate the sensor reading errors. The flow readings in industrial gas networks can have a constant bias in their measurement besides random noise and outliers. This inherent bias is difficult to detect with data analysis because of the constant shift in the parameters' sensor reading data.

In practice, bias is detected by putting a different but precise sensor next to the biased one. In this way, we can approximate the accurate reading of the precise sensor, and detect the bias amount that the malfunctioning sensors have. For the case study

problem, we need to measure the bias in a similar way. Specifically, we subtract the faulty sensor readings of the sensor from the non-erroneous sensor readings of the precise sensor. Then, we take the average of the difference over the selected period.

$$\tilde{B}_i = \text{mean}(\tilde{D}_i^t - D_i^t)$$

where $\tilde{D}_i^t = D_i^t + N_i^t + B_i$ after the outliers are eliminated. To do this, we simulate the actual demand data for the new sensor, compute $D_i^t + N_i^t$, subtract it from the biased and noisy sensor readings $\tilde{D}_i^t$ and take the average of the values for selected time horizon (n simulation points). The approximated bias values are the following: $\tilde{B}_i = [1.98, 3.97, 5.81]$ under the assumption that the AR model coefficients are known in demand models (4.4), (4.5), (4.6).

Recall from Equation (4.7) that the bias $B = \{2, 4, 6\}$ was added to the demand profiles. According to these results, the bias approximation $\tilde{B}$ for the three customer flow sensors accurately predicts the actual bias values B with no more than a 3.1% error.

Noise Filtering Approach

After eliminating the other sensor errors; that is, outliers and bias, we need to remove the noise from sensor readings which is the third and last step in the heuristic approach for sensor reading error elimination. To do so, we present a naive filtering algorithm (cf., Algorithm 4) below. This filtering approach is similar to both the well-known Savitzky-Golay and moving average filters (cf., Guiñón et al. 2007, Schafer 2011). While the proposed filter tries to separate the sensor noise from the process noise, which is not a simple task, it behaves conservatively by considering a portion of the actual demand volatility as sensor noise. However, this does not degrade the filter's performance and approximates the true demand signal as it can be seen from

Figures 4.9a, 4.9b, and 4.9c.

The filtering process starts by selecting the first observation as the starting observation. Then, it looks at the adjacent observation and decides whether the next observation differs significantly from the selected point based on the variance information of the noise. If the difference is significant, then we leave the next observation as it is. Otherwise, we take the moving average of the observations and replace the nominal values with the averaged values. The details of the noise filtering algorithm are given in Algorithm 4.

Algorithm 4 Noise filtering algorithm

Input: D

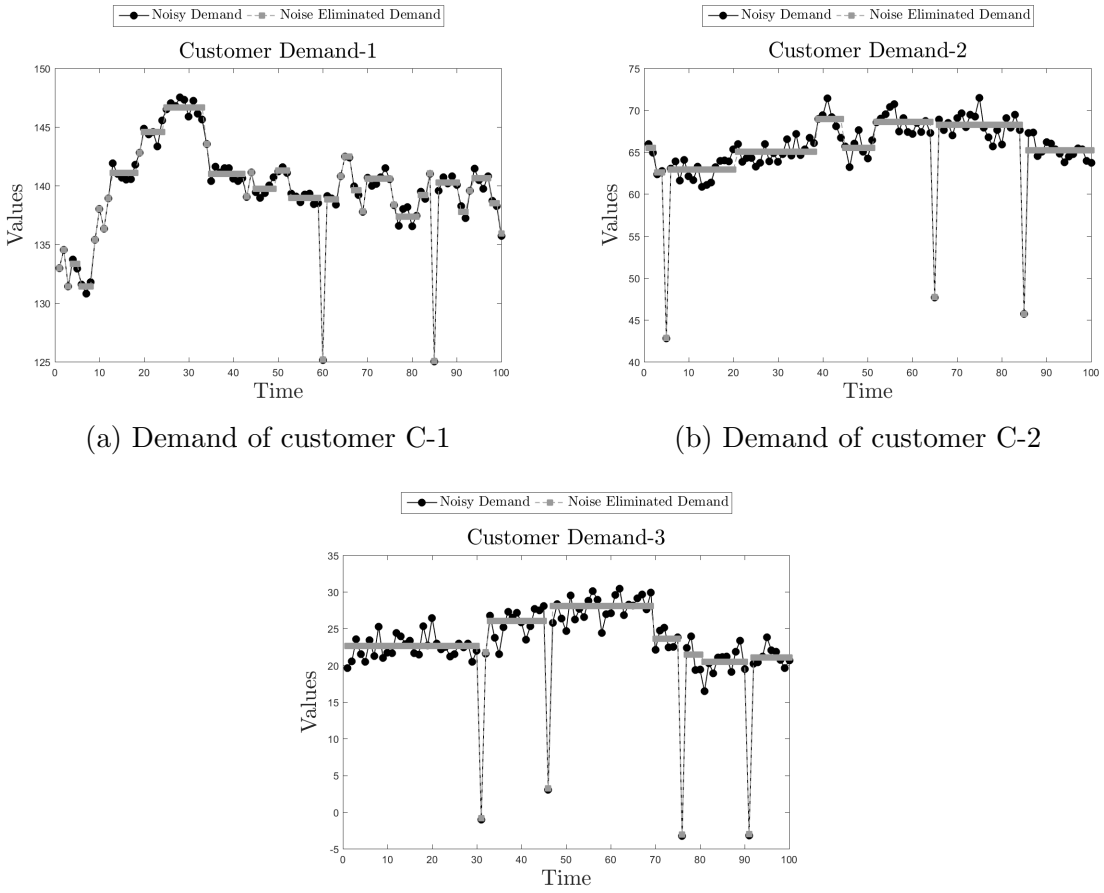
procedure NoiseFiltering
  $size \leftarrow$ length of D
  $D_{new}(1) \leftarrow D(1)$
  $k \leftarrow 2$
  for j=2:$size$ **do**
    $diff \leftarrow |D(j) - D_{new}(j-1)|$
    if $diff <= 3 * sigma$ **then**
      $avg = (\sum_1^k D(j-k+1))/k$
      $[D_{new}(j-k+1), \ldots, D_{new}(j)] = avg$
      $k \leftarrow k+1$.
    else
      $D_{new}(j) \leftarrow D(j)$
      $k \leftarrow 2$
    end if
  end for
end procedure

Output: D_{new}

In Figures 4.9a, 4.9b and 4.9c, we display the simulated data for customer demands for our problem setting before and after the noise filtering.

4.4.5 Numerical Results and Verification of the Methodology

Having the perfect knowledge of both the output of the optimization model (4.3) and the error-free demand profiles in Figure 4.4, gives us a chance to confirm our



(a) Demand of customer C-1

(b) Demand of customer C-2

(c) Demand of customer C-3

Figure 4.9: Noisy data and filtered data for customer demands.

methodology's validity. We already have the original production costs values, which we got from the optimization model by running it with faulty sensor demand readings ($\tilde{D}_i$). To compute the proposed KPIs, we calculate the average production costs values, and the production costs volatility, which is the variance of the values over the selected time horizon. After that, we select a sensor, and we use the perfect error knowledge that we created in Section 4.4.2 to eliminate its inherent sensor errors. While doing that, we keep the other sensors faulty to see the affect of elimination of a single sensor reading errors on the output. Then, we integrate the correct error free signal values for selected input parameter to the optimization model and reproduce the production costs values. Finally, we calculate the average production costs values,

and the production costs volatility by using the regenerated production values.

When we introduced sensor readings for customer demands ($\tilde{D}_i$) back in Section 4.4.2, we designed the errors to be the greatest for customer C-3, and the least for customer C-1. Now, we are going to inspect the analysis results to see what sensors need to be prioritized for the improvement.

Remember from Section 4.3 that we have two different KPIs. First, we check the results for KPI-1, which is the average change in the production costs' value over a time horizon. Table 4.3 displays the impact of the sensor reading error elimination on the production costs in a controlled way by using the optimization model and the simulated error information.

Fixed Sensors	Avg. Cost Function Value	Avg. Saving	Avg. Improvement
All Errors Present	$\$9.90E+04$	-	-
Fix Sensor-1 error only	$\$9.56E+04$	\$3400	3.45%
Fix Sensor-2 error only	$\$9.63E+04$	\$2700	2.69%
Fix Sensor-3 error only	$\$9.84E+04$	\$600	0.58%

Table 4.3: Error elimination of sensors with perfect knowledge (KPI-1)

After that, we apply our methodology to sensor readings as a univariate analysis to validate our methodology, and approximate the true signal of the selected input parameter. While doing that, we keep the other sensors faulty to see the affect of elimination of a single sensor reading errors on the output. Then, we integrate the approximated error free signal values for the selected input parameter to the predictive model, and reproduce the production costs values. Finally, we calculate the average production costs values, and the production costs volatility by using the regenerated production values.

Table 4.4 displays the impact of the sensor reading error elimination on the production costs value (KPI-1) with the sensor error elimination methodology given in Figure 4.2.

Fixed Sensors	Avg. Cost Function Value	Avg. Saving	Avg. Improvement
All Errors Present	$\$9.90E+04$	-	-
Fix Sensor-1 error only	$\$9.56E+04$	\$3400	3.45%
Fix Sensor-2 error only	$\$9.60E+04$	\$3000	3.07%
Fix Sensor-3 error only	$\$9.75E+04$	\$1500	1.54%

Table 4.4: Error elimination of sensors with the approximation methodology (KPI-1)

According to the comparison of results between Table 4.3 and Table 4.4, it follows that customer C-1 has the least amount of sensor reading errors, but it has the top priority to be fixed. This is because of two reasons: First, customer C-1 is using the most expensive gas in the network produced due to restrictions in the pipeline. Second, customer C-1's average demand value is much higher than the other customer demand profiles. These reasonings can be verified by the provided regression model's coefficients (4.8). The D_1's coefficient is reflecting the significance of the cost related to customer C-1. Although the results in Table 4.4 are not numerically precise to estimate the real changes given in Table 4.3, the difference between the numerical results are small, and produce a priority ranking in which the sensors are ordered as customer C-1, customer C-2 and customer C-3 in agreement with Table 4.3.

Secondly, we inspect the results for the second criteria we have: KPI-2, which measures the change in production costs' volatility. Table 4.5 displays the impact of the sensor reading error elimination on the production costs volatility in a controlled way by using the optimization model and the error information.

Fixed Sensors	Cost Function's Variance	Avg. Improvement
All Noise Present	5.52E+07	-
Fix Sensor-1 error only	5.14E+07	6.92%
Fix Sensor-2 error only	5.25E+07	5.05%
Fix Sensor-3 error only	5.42E+07	1.89%

Table 4.5: Error elimination of sensors with perfect knowledge (KPI-2)

Table 4.6 displays the impact of the sensor reading error elimination on the production costs volatility (KPI-2) with the sensor error elimination methodology given in Figure 4.2.

Fixed Sensors	Cost Function's Variance	Avg. Improvement
All Noise Present	5.52E+07	-
Fix Sensor-1 error only	5.29E+07	4.26%
Fix Sensor-2 error only	5.38E+07	2.56%
Fix Sensor-3 error only	5.43E+07	1.71%

Table 4.6: Error elimination of sensors with the approximation methodology (KPI-2)

Once again, customer C-1's sensor ranks first in the priority list for the sensor improvement, while customer C-3's sensor ranks last according to Table 4.5 and Table 4.6. This is again because of the customer C-1's and plant P-1's influences on the system. Similarly as for the KPI-1 results, the results in Table 4.6 are not numerically precise to estimate the real changes given in Table 4.3. However, we have the same priority ranking of the sensor maintenance ordered as customer C-1, customer C-2 and customer C-3 according to the results in both tables.

According to these results for both KPIs, we clearly see that both measures are suggesting us, in order of importance, to upgrade sensor-1 (demand of customer C-1) first, then sensor-2 (demand of customer C-2), and then sensor-3 (demand of customer C-3) in the industrial gas system. This is contrary to the order of the sensors' input error magnitudes discussed at the beginning of Section 4.3. Thus, these results show that analyzing the effects of input errors on the output (production costs) helps us to take better decisions. The computing time of the methodology for the case study problem with 3 customers and 3 flow sensors takes 1.23 seconds.

4.5 Implementation to the real pipeline system

In this section, we discuss the results obtained by applying the methodology to the company's real industrial gas network. The methodology we use for the real pipeline system is slightly modified to capture the real network's properties. First of all, we use the weighted ridge linear regression approach to obtain the predictive model instead of the ordinary least squares approach used in Section 4.4.3. The ridge regression approach is chosen because it addresses multicollinearities by imposing a penalty on the size of coefficients. Multicollinearities are possible between the input parameters because of the number of sensors (over 400 sensors) in the real network. This large number of predictors creates low ratio of number of observations to number of variables. It is also selected to be a weighted model because based on the expert's experience, the most recent information of the system carries more explanation of this very dynamic system. The industrial gas network of interest is a real time optimization system, where the optimal plant production flows need to be updated frequently. Thus, we desire to have a suggestion mechanism running and reporting suspicious sensor readings on a daily basis. Thus, the predictive model and heuristic analysis are implemented by using the last 100 data points read from the optimization model.

The predictive model is based on the weights calculated using the euclidean distance between the latest observation and the other samples in the set. The ridge regression estimate $\hat{\beta}$ is defined as the value of β that minimizes

$$\min \sum_{i=1}^{T} w^i (y^i - x^i \beta)^2 + \lambda \sum_{i=1}^{k} \beta_j^2, \tag{4.10}$$

where λ is chosen based on a 5-fold cross-validation, and the weights are computed by the following equations;

$$d(\mathbf{x}^i, \mathbf{x}^T) = \sqrt{(x_1^i - x_1^T)^2 + (x_2^i - x_2^T)^2 + \ldots + (x_k^i - x_k^T)^2},$$

where $\mathbf{x}^i = \{x_1^i, x_2^i, \ldots, x_k^T\}$.

$$w^t = \frac{1}{d(\mathbf{x}^i, \mathbf{x}^T)} \tag{4.11}$$

$$w^T = 1. \tag{4.12}$$

That is, we assign a weight of 1 to the most recent data point and for the other samples a weight equivalent to the inverse ratio of the Euclidean distance to the most recent sample.

Another modification in the methodology is leaving the bias elimination as an option to the end user. This is to avoid an extra investment cost at this stage of the study because the bias in sensors are estimated by placing a highly precise sensor to measure the same signal as the imprecise sensor measures. The general implementation of the methodology is aimed to be a statistical analysis only, and whenever the end user has the estimated bias information, there is an option in the interface for the user to give that information as an input to the analysis.

There are a few more specific implementations designed to make the analysis more meaningful and user-friendly. Based on the selection of the time period for the analysis, the end user can track these daily reports before taking the final decision for the upgrading or maintenance decision for the sensors. The interface summarizes the daily reports for the selected time period and suggests the most important ones for a possible upgrade or maintenance based on the analysis we describe in Section 4.5.1. Also, highly improving the sensor precision can be very costly. Due to this fact, the end user is also given the option to see the effects of partial improvements such as eliminating 50%, 75%, or 100% of the sensor errors. The default value in the results presented in this section is 75%. This is typically the fraction of the improvement that the company considers.

In the real industrial gas network problem, the uncertain inputs are not only

customer demands as in the case study in Section 4.4. Thus, random input variables can have very different units. In this case, parameters' regression coefficients can be easily influenced by the units in which the variables are measured, for example, gallons, pascals, bars, or grams in the real pipeline. Therefore, they do not provide a very reliable measure of the relative importance of the input variables. So, to compare the relative importance of the input parameters, input and output variables need to be standardized before drawing conclusions from their predictive model coefficients (cf., Bring 1994).

The described methodology is implemented in `MATLAB` and the historical sampling information is drawn by `SQL` queries from the company's database. The results are transferred to the company's online environment on a daily basis.

4.5.1 Implementation Results

Next, we present some of the results obtained by applying the methodology discussed here in the real industrial gas supply network. To protect the company's intellectual property, the sensor details are not provided. Due to the same reason, the input data for the analysis is not provided, as well as the exact results of the analysis. Instead, we provide a relative improvement on production costs and its volatility to prioritize the sensors in the system.

The pilot implementation for the analysis is chosen to be run in the business days (22 days) of October 2015. Customer demands are measured by flow sensors, and Table 4.7 provides the mean value and standard deviation values of the readings of 20 of these sensors.

Additionally, pairwise correlations between the sensors' readings are shown by a heatmap in Figure 4.10 where the individual values of the correlation matrix are represented as colors. Just by considering the pairwise correlations between these readings, the multicollinearity between the sensor readings is quite apparent.

Sensor #	Mean	Standard deviation
Sensor-1	12.245	0.124
Sensor-2	19.101	6.228
Sensor-3	43.006	2.132
Sensor-4	63.736	1.489
Sensor-5	42.595	1.307
Sensor-6	100.603	0.496
Sensor-7	115.115	0.307
Sensor-8	27.214	1.536
Sensor-9	15.141	0.105
Sensor-10	6.697	0.800
Sensor-11	30.167	0.441
Sensor-12	3.040	0.281
Sensor-13	85.729	5.838
Sensor-14	101.628	4.635
Sensor-15	22.418	4.364
Sensor-16	85.477	3.144
Sensor-17	39.449	5.152
Sensor-18	63.932	1.346
Sensor-19	0.807	2.064
Sensor-20	31.591	3.714

Table 4.7: Statistical measures of the sensor readings

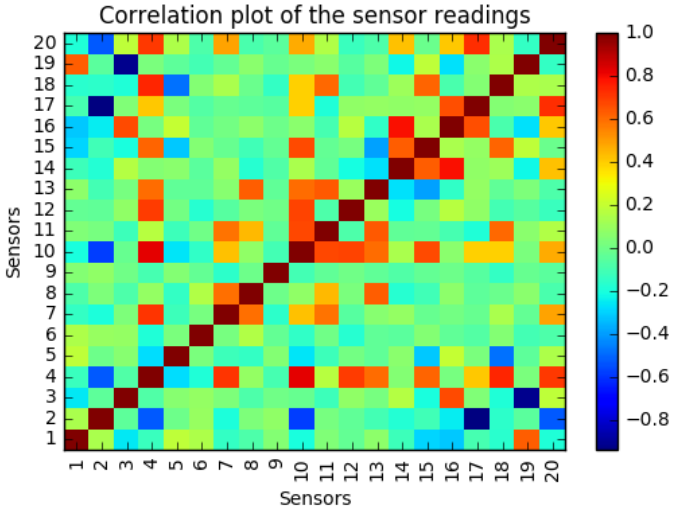


Figure 4.10: Pairwise correlations of the sensor readings.

The objective of the implementation is to run the univariate analysis daily with the updated results data coming from real-time optimization, and monitoring the parameter rankings based on their effect on the uncertainty and the value of the cost function. The resulting interface displays the sensor rankings in the order of positive savings on the production cost (KPI-1), and daily sensor rankings in the order of their sensor reading errors contribution on the objective function's volatility (KPI-2). While a single day analysis may not be meaningful due to the dynamic nature of real-time optimization system, a collection of the daily analysis can have strong suggestions to identify what sensors need to be upgraded or maintained.

Table 4.8 provides the top 5 sensors in this monthly collection, which ranked among the top 5 sensors based on KPI-1 in any of these daily runs. We note the number of times that any of these sensors were ranked in each of these top 5 positions. In each of the daily runs and in the summary, we only look at the top 5 sensors because based on our observations, most of volatility and cost improvements can be satisfied by upgrading or maintaining the top 5 sensors according to the daily analysis results.

According to Table 4.8's results, sensor-7 is leading the list by being ranked among the top-5 sensors 22 times out of 22 business days. It is ranked 5 times as the first sensor, 7 times as the 2nd sensor, and 9 times as the 3rd sensor, and only once as the 5th sensor in this set of analysis. sensor-8 also appears as one of the top sensors in almost the two third of the analysis' set.

	Rankings in KPI-1					
Parameter Names	1st	2nd	3rd	4th	5th	Total
Sensor-7	5	7	9	-	1	22
Sensor-8	3	4	1	5	1	14
Sensor-3	5	1	1	-	1	8
Sensor-11	2	1	-	1	3	7
Sensor-5	-	3	-	2	1	6

Table 4.8: Rankings of top 5 sensors based on reduction in production cost value (KPI-1)

The ranking of sensors displayed in Table 4.8 is one way of checking the significance of sensors in the system. We can also look at the average improvement in production cost's value over this month of analysis when the specified sensor is improved. These improvement values in percentage are calculated as follows

$$\text{Avg. Improvement} = \frac{(\hat{y} - \hat{y'})/\hat{y}}{T} * 100 \tag{4.13}$$

where $\hat{y}$ is the vector of the values of estimated objective functions before error elimination in the sensor, $\hat{y'}$ is the vector for the values of estimated objective functions after error elimination in the sensor, and T is the number of data points in a day. "Avg. Improvement" gives us the relative difference before and after the error elimination. Positive values for "Avg. Improvement" imply positive savings based on the univariate analysis for the selected sensor.

Parameter Names	Avg. Improvement in KPI-1
Sensor-7	1.415%
Sensor-8	0.266%
Sensor-3	0.095%
Sensor-5	0.065%
Sensor-11	0.040%

Table 4.9: Ranking of sensors based on average production savings (KPI-1)

The five highest priority sensors based on "Avg. Improvement" values in descending order are listed in Table 4.9. Table 4.9 suggests us that we would reduce the daily production costs by around 1.4% in average if we go ahead and fix the errors of sensor-7. sensor-7 is known as the flow sensor of an important customer in the system, and the sensor reading errors affect the production cost significantly. The daily average improvement is based on a month of analysis and it is an important number when it is translated to real money. For this reason, KPI-1 is definitely an important indicator that upper management would certainly consider.

The numbers in Table 4.9 are not additive, in other words, one cannot guarantee the amount of improvement would be equal to the summation of percentage values if multiple sensors are decided to be upgraded. This is due to the univariate nature of the analysis, i.e. we approximate the true signal for only one selected sensor at a time while keeping the other sensor reading errors present in the system.

Also, although the rankings and/or identities of the sensors in Table 4.8 and Table 4.9 are almost identical to each other in this analysis, they do not necessarily have to match with each other as they rank the sensors based on different evaluations for the same key performance index. However, the similarity between the order of the sensors in Table 4.8 and Table 4.9 is a strong indicator to consider prioritizing the suggested sensors for the maintenance based on the selected KPI. This also applies to the results going to be presented in Table 4.10 and Table 4.11 based on KPI-2.

As we discussed in Section 4.1, it is also important to consider KPI-2 as a decision criterion while selecting the sensors for maintenance. Table 4.10 provides a summary of the analysis realized in October 2015 for the key performance indicator KPI-2. Specifically, it provides the number of times that a sensor is ranked among the top 5 sensors in terms of KPI-2 in any of these daily runs. Similar to the results of the KPI-1 analysis, sensor-7 was ranked as the first sensor for a possible precision upgrade in 12 out of 22 business days. On the other hand, it was ranked 5 times as the second sensor and 4 times as the third sensor to be improved in this monthly analysis.

	Rankings in KPI-2					
Parameter Names	1st	2nd	3rd	4th	5th	Total
Sensor-7	12	5	4	-	-	21
Sensor-3	5	4	3	2	2	16
Sensor-6	1	2	3	3	2	11
Sensor-11	-	5	1	1	1	8
Sensor-9	1	5	-	1	1	8

Table 4.10: Rankings of top 5 sensors based on volatility reduction in production costs (KPI-2).

Similar to Table 4.8, one can look at the average improvement in production cost's volatility over the selected time period when the specified sensor is improved. These improvement values in percentage are calculated as follows

$$\text{Improvement} = \frac{\text{Var}(\hat{y}) - \text{Var}(\hat{y'})}{\text{Var}(\hat{y})} * 100 \tag{4.14}$$

where $\text{Var}(\hat{y})$ is the variance of computed objective function values before upgrading the sensor, and $\text{Var}(\hat{y'})$ is the variance of estimated objective function values after upgrading the sensor. "Improvement" gives us the relative difference before and after the operation. Positive values for "Improvement" imply reduction in the objective function's volatility.

The top five sensors based on the average volatility reduction in the case of their maintenance are listed in Table 4.11. According to these results, volatility in production costs reduces around 2.9% in daily average when sensor-7's reading errors are tackled. Although the relationship is not obvious, due to the implied costs and inconveniences of having uncertainties in the problem, the reduction in production cost's volatility could result in great savings in the production cost's value. Similar to the results displayed in Table 4.9, the given percentage improvement for the sensors in Table 4.11 are not additive.

Parameter Names	Avg. Improvement in KPI-2
Sensor-7	2.910%
Sensor-3	1.922%
Sensor-6	0.765%
Sensor-8	0.399%
Sensor-11	0.341%

Table 4.11: Ranking of sensors based on average volatility reduction (KPI-2)

After reviewing the results of the analysis, the end user can decide what time period he or she is concerned with and what criteria are important for the company

for the sensor improvements. In this pilot analysis summary conducted with October 2015's optimization model's historical data, sensor-7's reading errors significantly dominate the system in comparison with the other sensors in the system according to both of these KPI's and their ranking and/or average improvement criteria.

Finally, since the methodology is designed as a heuristic approach, it is crucial to report the computational run times of the methodology for different number of sensors. The methodology is run 10 times for the instances with different number of sensors. Table 4.12 presents the average run time and the standard deviation of the time needed to run these experiments. Although the time required to solve an instance typically increases with the number of sensors, this solution time depends on the number of outliers, making experiments with lower number of sensors take longer in average. For example, this is the case for the cases with 350 and 400 sensors. In general, computation times are small and in the range of 10 seconds.

Number of sensors	Average run time (sec.)	Standard deviation of run times (sec.)
400	9.111	6.639
350	12.620	13.991
300	7.735	5.699
250	3.776	0.858
200	3.532	3.985
150	2.377	1.124
100	2.622	1.418

Table 4.12: Summary of 10 computational run times of heuristic approach for different number of sensors

On the other hand, the propietary optimization of the real system is inflexible and cannot be run for different network sizes. This is due to the nature of the model which is highly complex and nonadjustable. We refer the reader to van den Heever and Grossmann (2003) for some recorded times of a similar optimization model on networks of different sizes.

4.6 Conclusion

In this paper, we presented a practical sensor fault identification and improvement methodology for an industrial gas supply network based on sensitivity and data analysis. We constructed predictive models based on global sensitivity analysis tools and then used some data analysis techniques to approach sensor's error-free signals. Then, we analyzed the benefits of having an error free signal for each specified sensor. To validate the methodology, we presented the application of the methodology in a simple case study problem. Then, with a few modifications, we extended the same methodology to the real industrial gas network system. The verified approximation gives us the necessary tools to reduce the measurement inefficiencies in the network.

The results of the analysis are currently being used in decision-making processes to detect which sensors are providing suspect readings in a given period of time. Based on the application of the analysis, sensor repairs are going to be selected and realized based on the cost savings and/or the reduction of the volatility of the production cost for the company.

Chapter 5

Clustering in Portfolio Optimization

In early 2000's, it was shown that the classical mean variance (MV) portfolio allocation model and some of it's extensions, designed to reduce the estimation error, were not able to outperform the naive equally weighted portfolio allocation strategy. Since then, there are new allocation strategies and filtering techniques that have been proposed based on hierarchical clustering to obtain better performing portfolios when tested with out of sample data. The integration of the hierarchical clustering approach to the portfolio allocation decisions has been shown to be effective under appropriate circumstances in the relevant literature. In this study, we compare these relatively recent portfolio allocation techniques with other well known asset allocation techniques in a single, and widely accepted experimental design setting. While our results confirm some of the conclusions already drawn in the literature, they contradict with some relevant ones, which leads to new insights regarding the performance of these different portfolio allocation strategies.

5.1 Introduction

The portfolio optimization is one of the most studied topics in quantitative finance (see, e.g. Elton et al. 2009). The widely used sample-based mean-variance framework of Markowitz (Markowitz 1952) uses the first and second moment estimates of financial asset returns, to construct mean-variance efficient portfolio. Although the mean-variance model is highly regarded for its theoretical & practical properties, it is known to be highly sensitive to estimation errors (see, e.g. DeMiguel and Nogales 2009, Michaud 1989, Chopra et al. 2011).

In the literature, a considerable effort has been invested to reduce the estimation error of the sample moments which highly affect the out of sample performance of mean-variance portfolios. Among these methods are the well known bayesian approach (see, e.g. Bawa et al. 1979), shrinkage estimators (see, e.g. Jobson et al. 1979), robust portfolio allocation rules (see, e.g. Goldfarb and Iyengar 2003), methods that focus on reducing the error in estimating the covariance matrix (see, e.g. Ledoit and Wolf 2004, Conlon et al. 2007, Pantaleo et al. 2011) and portfolio rules that impose shortselling constraints (see, e.g. Jagannathan and Ma 2003). In a comprehensive study, DeMiguel and Nogales (2009) compared the out of sample performance of a number of these methodologies with the naive equally weighted portfolio diversification rule, which allocates a $\frac{1}{N}$ weight to each of the N assets in the portfolio. Here, the equally weighted allocation rule is chosen as the benchmark strategy because it does not involve any estimation error and it is simple to use for an investor. In their work, DeMiguel and Nogales (2009) show that out of the models evaluated, none performs consistently better than the equally weighted portfolio policy in terms of 3 performance measures: Sharpe ratio, certainty-equivalent return and turnover.

While a wide range of models were compared to the equally weighted allocation policy in DeMiguel and Nogales (2009)'s work, recent ideas to reduce the error in

estimating the covariance matrix were not investigated. One of these methods uses hierarchical clustering to filter the covariance matrix and improve the out of sample performance of mean variance portfolios. Moreover, there are new methodologies in the literature that use clustering as a subroutine to obtain novel portfolio allocation strategies (see, e.g. López de Prado 2016, Raffinot 2016). Our objective in this study is to test recent hierarchical clustering based portfolio allocation strategies in the widely regarded experimental design proposed in DeMiguel and Nogales (2009)'s study, and compare the out of sample performances of these strategies with the naive equally weighted strategy and sample based mean variance strategies and its extensions.

Clustering financial time series and the study of correlation networks first appears in the seminal work of Mantegna (1999). Clusters are obtained based on correlation coefficients between the financial assets and they provide a unique indexed hierarchy between the assets. This unique hierarchy can be represented by a minimum spanning tree which also corresponds to the dendrogram obtained using the Single Linkage Hierarchical Clustering Algorithm (Gower and Ross 1969). The resulting cluster information is used to apply either as a filtering procedure for the covariance matrix (Tumminello et al. 2010) or they help practitioners to develop financial applications such as portfolio allocation strategies (López de Prado 2016, Raffinot 2016) and financial policy makings (Harmon et al. 2010).

The use of a filtered correlation matrix in the mean-variance portfolio selection framework has been considered by Pantaleo et al. (2011) and Tola et al. (2008). Their main conclusion is that the use of filtered covariance matrices based on clustering are particularly useful in mean-variance portfolio optimization with short sales. When short sales are not allowed, the filtering procedures, including clustering based filtering, are unable to outperform the sample-based mean-variance model. Similar results also apply to the tests with the sample-based minimum variance model with short sale constraints. However, we have not encountered any study in the literature comparing

the out of sample performance of the clustering based filtering approaches with the naive equally weighted portfolio allocation strategy.

In another study, López de Prado (2016) uses clusters to reorganize the covariance matrix and applies a portfolio allocation strategy called "Hierarchical Risk Parity". Here, the objective is to use hierarchical clustering to identify the hierarchical structure between the financial assets. Then the covariance matrix based on the cluster information and inverse-variance allocation strategy to determine the weights (López de Prado 2016). His preliminary results on the simulated dataset are promising. However, to the best of our knowledge, there is no comprehensive empirical investigation comparing this recent methodology to the classical portfolio allocation strategies.

Here we investigate the out of sample performance of clustering based filtering methodologies and the "Hierarchical Risk Parity" methodology, and compare them with the performance of the traditional risk allocation strategies such as the sample-based mean-variance model with short sale constrains, the sample-based minimum-variance model with short sale constraints, the equally-weighted portfolio allocation method, and the traditional risk parity method.

In our empirical study, we do not investigate the unconstrained models for sample-based mean-variance and sample-based minimum-variance strategies because it is already known that different filtering procedures of covariance matrix, including hierarchical clustering, yield significantly improved out-of-sample risk and Sharpe ratio results than the unconstrained models (see, e.g. Jagannathan and Ma 2003, DeMiguel and Nogales 2009, Pantaleo et al. 2011, Tola et al. 2008). According to Jagannathan and Ma (2003) imposing short sale constraints on the sample-based models improves the performance in the same way that shrinking the expected return towards the average does. Similarly, imposing a shortsale constraint on the sample-based minimum-variance model is equivalent to shrinking elements of covariance matrix. Although, it is already shown in Pantaleo et al. (2011) that imposing a short sale constraint

on the sample based mean-variance model and its extensions produce a competitive out-of-sample risk performance compared to using hierarchical clustering to filter the covariance matrix in the mean-variance model and its extensions, the out-of-sample Sharpe ratio performance of the models incorporating short sale constraints with clustering based filtering were not investigated. In this work, we address this with an empirical investigation.

Our results confirm many of the conclusions drawn in the literature. First, we replicated and verified some of the results reported in DeMiguel and Nogales (2009)'s work. In particular, for the Kenneth French's ten industry portfolios and the US equity market portfolio dataset (French 2017), our experimental setting returned almost exactly the same values reported in DeMiguel and Nogales (2009) for three allocation methods; namely, mean variance model with shortsale constraints, minimum-variance model with shortsale constraints and the equally-weighted portfolio. Second, we show that the hierarchical clustering based filtering methods to reduce the estimation error in the covariance matrix do not improve the out of sample risk or Sharpe ratio of mean-variance models with shortsale constraints. Third, López de Prado (2016) reported favorable results for the Hierarchical Risk Parity method compared to the minimum-variance model with shortsale constraints and the traditional risk parity model. Although this recent method results in highly diversified portfolios, it does not consistently outperform the traditional portfolio allocation strategies and it's performance is particularly comparable to the traditional risk parity method.

The chapter is organized as follows. In Section 5.2, we briefly describe the portfolio allocation methodologies investigated in this study. In Section 5.3, we present the datasets and the experimental design to compare different portfolios. We provide the numerical results in Section 5.4. Finally, Section 5.5 concludes the chapter.

5.2 Description of the Portfolio-Allocation Models Considered

In this section, we will briefly describe the methodologies considered here. We investigate the out of sample performances of Equally Weighted allocation (EW), sample-based Mean-Variance allocation with short sales constraints (MV-c), Mean-Variance allocation with short sales constraints and with cluster-based filtering (MV-SLCA-c), sample-based Minimum-Variance allocation with short sales (Min-c), Minimum-Variance allocation with short sales and with cluster based filtering (Min-SLCA-c), Traditional Risk Parity Allocation exemplified by the Inverse-Variance Portfolio (IVP) as it is described in López de Prado (2016), the recently proposed Hierarchical Risk Parity (HRP) method by López de Prado (2016), and Mean-Variance allocation models with short sales and benchmark return constraints (MV-c-RetHRP) and it's clustering based filtering version (MV-c-SLCA-RetHRP), where the benchmark return is obtained by the sample average return of HRP portfolio allocation.

5.2.1 Equally-Weighted Allocation

The equally-weighted allocation strategy is to hold a portfolio where each of the N assets has a weight of $\frac{1}{N}$. This strategy does not involve any parameter estimation or optimization, so it is estimation error free. This strategy is chosen as the benchmark strategy to compare the out-of-sample Sharpe ratios and turnover values of the strategies investigated.

5.2.2 Mean-Variance Portfolio Optimization

Given N risky assets, the mean variance portfolio is the solution to the optimization problem

$$\begin{aligned} \min \quad & w^{\mathrm{T}}\hat{\Sigma}w - \tfrac{1}{\gamma}w^{\mathrm{T}}\hat{\mu} \\ \text{s.t.} \quad & w^{\mathrm{T}}e = 1 \\ & w \geq 0 \end{aligned} \tag{5.1}$$

,

where $w \in \mathbb{R}_+^N$ is the vector of portfolio weights, $w^{\mathrm{T}}\hat{\mu}$ is the sample mean of the portfolio returns, and $w^{\mathrm{T}}\hat{\Sigma}w$ is the sample variance of the portfolio returns, with $\hat{\Sigma}$ denoting the sample covariance matrix of the asset returns. The constraint $w^{\mathrm{T}}e = 1$, where $e \in \mathbb{R}^N$ is the vector of all ones, ensures that the portfolio weights sum to one, and the constraint $w \geq 0$ ensures that there is no short-selling.

For different values of the risk aversion parameter γ, mean-variance portfolios on the efficient frontier are obtained. In our experimental design, γ is set to 1 following DeMiguel and Nogales (2009). The MV-c and MV-SLCA-c methods are based on the model described above. MV-c uses the sample covariance matrix $\hat{\Sigma}$ for the covariance matrix estimate. MV-SLCA-c replaces the sample covariance matrix $\hat{\Sigma}$ in Model (5.1) with the filtered covariance matrix $\hat{\Sigma}_{SLCA}$ obtained using hierarchical clustering in the model. The procedure to obtain $\hat{\Sigma}_{SLCA}$ is discussed in Section 5.2.4.

5.2.3 Minimum-Variance Portfolio Optimization

The minimum-variance portfolio is the mean-variance portfolio corresponding to the highest risk aversion parameter ($\gamma = 0$). Thus, it can be computed by solving the following optimization problem:

$$\begin{aligned} \min \quad & w^{\mathrm{T}}\hat{\Sigma}w \\ & w^{\mathrm{T}}e = 1 \\ & w \geq 0 \end{aligned} \tag{5.2}$$

Notice that the estimation errors of the expected asset returns does not affect the minimum variance portfolio, since the expected asset returns are not involved in the model.

Min-c and Min-SLCA-c methods are based on the model described above. Min-c uses the sample covariance matrix $\hat{\Sigma}$ as the covariance matrix estimate. Min-SLCA-c replaces the sample covariance matrix $\hat{\Sigma}$ in Model (5.2) with the filtered covariance matrix $\hat{\Sigma}_{SLCA}$ obtained using hierarchical clustering.

5.2.4 Hierarchical Clustering Based Filtering

Hierarchical clustering methods are clustering procedures in which elements are iteratively merged together in clusters of increasing size according to their degree of similarity. In his seminal paper, Mantegna (1999) investigates the correlation coefficient matrix to detect the hierarchical structure present in a portfolio of N assets traded in a financial market. We describe his methodology next.

Given N financial assets and historical returns of the financial assets, one can calculate the correlation coefficients $\rho_{i,j}$ between the assets i and j as a similarity measure between pairs of assets i, j for all $i, j = 1, \ldots, n$. At the beginning of the clustering procedure, each asset defines its own cluster. Then, at any iteration, the two clusters with the largest correlation are merged together in a single cluster and the cluster set is updated. At the second and further iterations different similarities between clusters can be defined, each one characterizing a specific hierarchical clustering procedure. The number of total clustering iterations is $N-1$. The hierarchical clustering procedure considered in this study is called Single Linkage Clustering Algorithm (SLCA). In fact the SLCA uses the maximal correlation coefficient between distinct groups of elements as the similarity measure between clusters. If a new cluster q is formed from clusters h and k, then the similarity between cluster q and any

other cluster j is given by

$$\rho_{qj} = \max\{\rho_{hj}, \rho_{kj}\}$$

indicating that the similarity between any element of cluster q and any element of cluster j is the similarity between the two most similar entities in clusters q and j.

While The SLCA algorithm leads to a single cluster eventually, it determines a minimal spanning tree connecting the n stocks of the portfolio with $n-1$ links (Gower and Ross 1969). This tree is a dendrogram where each node α_k is associated with the distance ρ_{α_k} (similarity measure) between the two clusters of elements merging together in the node α_k. The dendrogram stores the key information of this procedure, such as how the clusters are historically formed and the similarity distance between each other.

One can construct a filtered correlation matrix C_{SLCA} from the resulting dendrogram from the SLCA algorithm. C_{SLCA} has $n-1$ distinct correlation coefficient values instead of $n(n-1)/2$ distinct elements characterizing the sample covariance matrix, and it is shown that C_{SLCA} is an ultrametric correlation matrix which is always positive definite if all the elements of the matrix are positive (Tumminello et al. 2010, Marti et al. 2017). This condition is required to efficiently solve Problems (5.1) or (5.2). Once C_{SLCA} is constructed, the covariance matrix estimate Σ_{SLCA} can be obtained by multiplying the entries of C_{SLCA} by the assets' sample standard deviations (Pantaleo et al. 2011).

5.2.5 Hierarchical Risk Parity

More recently, López de Prado (2016) introduced a new portfolio allocation methodology called Hierarchical Risk Parity (HRP). This strategy is based on the cluster information resulting from the single linkage hierarchical clustering algorithm and the inverse-variance allocation of the weights. López de Prado (2016) points out that

correlation matrices lack the notion of hierarchy and make no difference between assets, which allows weights to vary freely in unintended ways. However, in reality not every financial asset is actually a substitute of each other. While some of them are closer substitutes of one another, some other assets are complimentary to one another.

In the first part of the HRP methodology, the single linkage hierarchical clustering based on the sample correlation matrices is used to find the dendrogram representing the hierarchical structure as it is described in Section 5.2.4. Once the clusters have been determined, the capital should be efficiently allocated both within and across groups. For that purpose, López de Prado (2016) reorganizes the rows and columns of the covariance matrix based on the cluster information, so that the largest values lie along the diagonal. This quasi-diagonalization of the covariance matrix provides a useful property: similar investments are placed together, and dissimilar investments are placed far apart. Finally, the weights are assigned to the assets by the inverse-variance allocation technique which is optimal for a diagonal covariance matrix.

5.2.6 Traditional Risk Parity

Traditional risk parity allocation is represented by the inverse variance portfolio allocation (IVP) where the allocated weights are given by,

$$w_i = \frac{1/\sigma_i^2}{\sum_{i=1}^{N} 1/\sigma_i^2}, \quad i = 1, \ldots, n$$

The rationale behind the inverse variance allocation technique is to assign large weights to low volatile stocks.

5.2.7 Mean-Variance Portfolio Optimization with a benchmark return constraint

Mean-variance models can be formulated in multiple ways (Braga 2016). We provided the one with the risk aversion parameter in Model (5.1). To test whether mean-variance formulation could return the lower risk while achieving the same portfolio return achieved by HRP strategy, we reformulate the Model (5.1) with a benchmark return constraint and we define the benchmark return value as the HRP portfolio return given by the following equation

$$\mu_{(\text{HRP})} = w_{(\text{HRP})}^{\text{T}} \mu$$

where w_{HRP} are the weights obtained by HRP allocation and μ is the sample average of the asset returns. The mean-variance model with the benchmark return constraint then can be formulated by following Model (5.3)

$$\begin{aligned} \min \quad & w^{\text{T}} \hat{\Sigma} w \\ \text{s.t.} \quad & w^{\text{T}} \mu = \mu_{(\text{HRP})} \\ & w^{\text{T}} e = 1 \\ & w \geq 0 \end{aligned} \tag{5.3}$$

,

5.3 Experimental Setting

5.3.1 Description of Empirical Datasets

10+1 Industry Portfolios

The $N = 10{+}1$ Industry portfolios dataset from Ken French's financial database (French 2017) is one of the datasets that DeMiguel and Nogales (2009) used in their exper-

iments. This dataset is chosen to replicate some of the results that provided in his study. The dataset consists of monthly excess returns of 10 industry portfolios in the United States. The 10 industries considered are Consumer-Discretionary, Consumer-Staples, Manufacturing, Energy, High- Tech, Telecommunication, Wholesale and Retail, Health, Utilities, and Others. The monthly returns range from July 1963 to November 2004. The dataset is augmented by adding as an asset the excess return on the US equity market portfolio (MKT).

30 Industry Portfolios

The dataset consists of monthly excess returns on 30 industry portfolios in the United States. The monthly returns range from July 1963 to February 2017 and were obtained from Ken French's financial database (French 2017).

SPI sectors

The "SPI sectors" dataset consists of daily returns on 10 value weighted industry portfolios formed by using the Global Industry Classiffication (GICS) developed by Standard & Poor's. The 10 industries considered are Energy, Material, Industrials, Consumer-Discretionary, Consumer-Staples, Healthcare, Financials, Information-Technology, Telecommunications, and Utilities. The data span from January 1995 to August 2016. This dataset is obtained from Thomas Raffinot (Raffinot 2016).

Dow Jones Industrial Average

The dataset consists of daily excess returns on 28 Dow Jones industrial average index companies in the United States (the stocks in Dow Jones 30 except for Visa and Goldman Sachs stocks). The daily returns range from January 1996 to August 2016 and this dataset is created from the dataset constructed by Thomas Raffinot for S&P-500 daily returns (Raffinot 2016).

5.3.2 Comparison Measures

Following DeMiguel et al. (2009), the methodologies are tested on the given datasets based on a rolling-window approach. Given a T-month or T-day long dataset of returns, an estimation length M is chosen. To test the sensitivity of the input parameters, different estimation lengths are considered. For the datasets involving daily returns, specifically, the values $M = [66, 130, 260, 520]$ days are used. For the datasets involving monthly returns, we tried $M = [30, 60, 120]$ months. The portfolios are rebalanced every R periods. The common practice in the industry is rebalancing the portfolio every month. Thus, we use $R = 1$ month for the datasets involving monthly returns, and $R = 22$ days for the datasets involving daily returns. In each rebalancing period t, starting from $t = M+1$, parameters for the particular strategy are estimated using the M time units preceding t. The weights are calculated by the chosen strategy and used to compute the return in the next period $t+R$. This process is continued by adding the R returns from the next period in the dataset and dropping the earliest R returns, until the end of the dataset. The outcome of this rolling-window approach is a vector of $\frac{T-M}{R}$ monthly out of sample returns generated by each of the portfolio methods described in Section 5.2 for each of the datasets described in Section 5.3.1.

Given the time series of monthly out-of-sample returns generated by each strategy in each dataset, several comparison criteria are computed:

Risk

Out-of-sample risk of strategy k is estimated by the sample variance of the out-of-sample excess monthly returns $\hat{\sigma}_k$. To test whether the out of sample risk of two strategies are statistically distinguishable, we also compute the p-value by conducting an F-test for the null hypothesis that normally distributed iid samples from two populations have the same variance.

Sharpe ratio

Out-of-sample Sharpe ratio of strategy k, defined as the sample mean of out-of-sample excess returns (over the risk-free asset), $\hat{\mu}_k$, divided by their sample standard deviation, $\hat{\sigma}_k$. To test whether the Sharpe ratios of two strategies are statistically distinguishable, we also compute the p-value of the difference, using the approach introduced by (Jobson and Korkie 1981) and referenced in (DeMiguel et al. 2009). Specifically, given two portfolios i and j, with $\hat{\mu}_i, \hat{\mu}_j, \hat{\sigma}_i, \hat{\sigma}_j, \hat{\sigma}_{i,j}$ as their estimated means, variances, and covariances over a sample of size $\frac{T-M}{R}$, the test of the hypothesis $H_0 : \frac{\hat{\mu}_i}{\hat{\sigma}_i} - \frac{\hat{\mu}_j}{\hat{\sigma}_j} = 0$ is obtained via the test statistic

$$\hat{z}_{JK} : \frac{\hat{\mu}_i * \hat{\sigma}_j - \hat{\mu}_j * \hat{\sigma}_i}{\sqrt{\hat{\nu}}},$$

where

$$\hat{\nu} : \frac{1}{(T-M)/R}\left(2\hat{\sigma}_i^2\hat{\sigma}_j^2 - 2\hat{\sigma}_i\hat{\sigma}_j\hat{\sigma}_{i,j} + 0.5\hat{\mu}_i^2\hat{\sigma}_j^2 + 0.5\hat{\mu}_j^2\hat{\sigma}_i^2 - \frac{\hat{\mu}_i\hat{\mu}_j}{\hat{\sigma}_i\hat{\sigma}_j}\hat{\sigma}_{i,j}^2\right).$$

Turnover

To get a sense of the amount of trading required to implement for each portfolio strategy a third performance metric computed is the portfolio turnover for each period, defined as

$$TO_{kt} = \sum_{j=1}^{N}\left(|\hat{w}_{k,j,t+R} - \hat{w}_{k,j,t^+}|\right)$$

where $\hat{w}_{k,j,t^+}$ is the portfolio weight before rebalancing at $t+R$ which is different than the values $\hat{w}_{k,j,t}$ or $\hat{w}_{k,j,t+R}$ due to changes in asset prices between t and $t+R$.

After finding the vector of turnover values for each strategy and for each rebalancing period, we compute the sample mean and the sample standard deviation of the vector of turnover values. The average of this vector of values can be interpreted as

the average percentage of wealth traded each period. To test whether the average percentage of wealth traded each period of two strategies are statistically distinguishable, we compute the p-value of the difference by applying a two-sample t-test.

5.4 Numerical Results

In this section, we present results obtained by the portfolio allocation strategies described in Section 5.2 for the datasets described in Section 5.3.1. For each strategy, we compute the realized out-of-sample risk, the Sharpe ratio and the turnover.

5.4.1 Risk

Portfolio riskiness is one of the comparison measures that we are interested in. The risk is calculated by the variance of the out-of-sample returns. We analyze the results based on the time unit of the returns in the datasets.

Figure 5.1 and Figure 5.2 show the realized risk for the portfolio strategies as a function of training periods for the datasets involving monthly returns and for the datasets involving daily returns respectively. We observe that for all the datasets, the mean-variance models result in comparably greater risk with both sample covariance matrices (MV-c) and filtered covariance matrices (MV-SLCA-c). On the other hand, all the other strategy's risk values are comparably closer to each other. Overall, Min-c, Min-c-SLCA, MV-c-RetHRP and MV-c-SLCA-RetHRP are the strategies returning the lowest risk values.

Moreover, using filtered covariance matrices instead of sample covariance matrix estimates in mean-variance (MV-c), minimum-variance (Min-c) and mean-variance with benchmark constrained models (MV-c-RetHRP) do not improve the portfolio risk significantly. This is a consequence of using short sale constraints in the models as it is previously stated in Section 5.1. On the other hand, in general, HRP portfolios'

risk is almost equivalent to IVP portfolios' risk and it can be concluded that the HRP portfolios are certainly not the best performing portfolios in terms of the out of sample risk for the investigated datesets and the portfolio allocation strategies.

To measure the statistical significance of the portfolios risk performances, we select the minimum variance strategy (Min-c) as the benchmark strategy since it is one of the strategies that yields the lowest out-of-sample risk compared to the other strategies in the experiments as it can be verified in Figures 5.1 and 5.2 . The numerical results and p-values, which are provided in Table 5.1 and Table 5.2 for all the investigated datasets and for the selected training periods, verify the conclusions reached from Figure 5.1 and Figure 5.2. MV-c and MV-c-SLCA portfolios have significantly greater risk than Min-c portfolios for all the datasets. EW, HRP and IVP strategies also perform significantly worse than the Min-c portfolio allocation strategy in general. For these three strategies, the difference between their portfolios' risk and the Min-c portfolio risk are quite significant for the dataset "30 Industry Portfolios" while it is the least significant for the dataset "SPI Sectors". Moreover, we find out that the portfolio risks for MV-SLCA-c, MV-c-RetHRP and MV-c-SLCA-RetHRP strategies do not differ significantly from the Min-c portfolio risk.

Finally, using HRP allocation's expected portfolio return as a benchmark parameter in benchmark-return-constrained mean-variance models (MV-c-RetHRP and MV-c-SLCA-RetHRP) generally yields better out-of-sample risk values compared to the mean-variance models using risk aversion parameter $\gamma = 1$ (MV-c and MV-c-SLCA). This is due to restricting the portfolio's expected return to a fix value reduces the effect of the estimation errors on the expected asset returns.

5.4.2 Sharpe ratios

The out-of-sample Sharpe ratio is another metric that we consider when comparing different portfolio allocation strategies. The Sharpe ratio is the risk adjusted returns

Figure 5.1: The realized risk for the portfolio strategies as a function of training periods for the datasets involving monthly returns of Industry Portfolios ($N = 10 + 1$) and Industry Portfolios ($N = 30$)

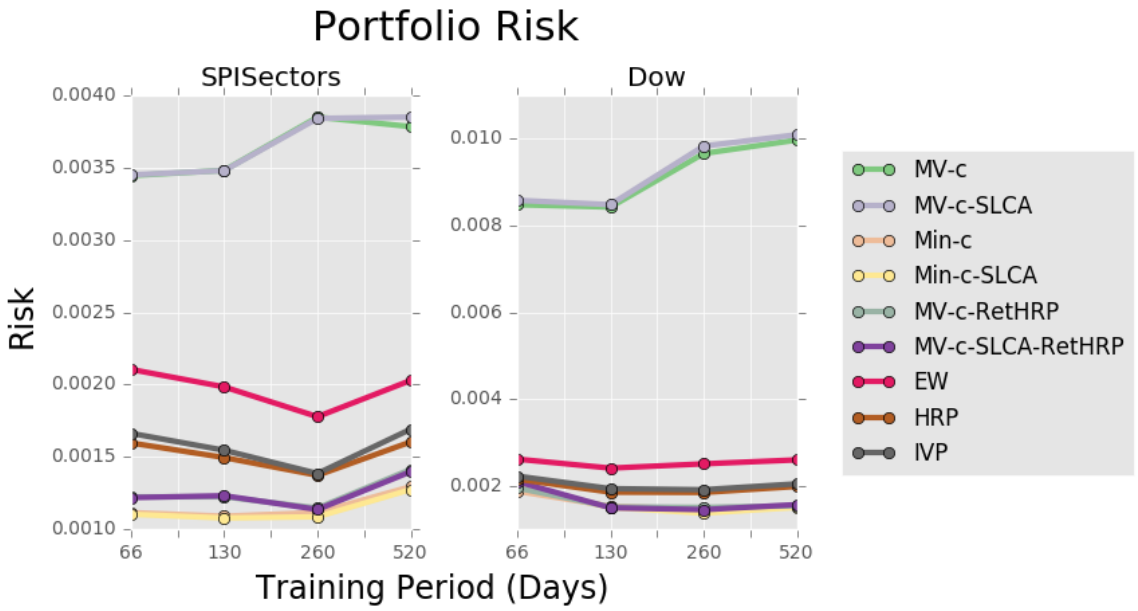


Figure 5.2: The realized risk for the portfolio strategies as a function of training periods for the datasets involving daily returns of SPI Sectors ($N = 10$) and Dow Jones stocks ($N = 28$)

	Industry Portfolios, $N = 10 + 1$				Industry Portfolios, $N = 30$			
	$M = 60$		$M = 120$		$M = 60$		$M = 120$	
Strategies	Risk	p	Risk	p	Risk	p	Risk	p
MV-c	0.0038	0.00	0.0040	0.00	0.0063	0.00	0.0068	0.00
MV-c-SLCA	0.0038	0.00	0.0042	0.00	0.0064	0.00	0.0071	0.00
Min-c(*)	0.0014	*	0.0014	*	0.0013	*	0.0013	*
Min-c-SLCA	0.0014	0.44	0.0014	0.40	0.0013	0.38	0.0013	0.48
MV-c-RetHRP	0.0015	0.33	0.0014	0.47	0.0014	0.25	0.0013	0.47
MV-c-SLCA-RetHRP	0.0014	0.42	0.0014	0.47	0.0013	0.41	0.0013	0.46
EW	0.0019	0.00	0.0019	0.00	0.0023	0.00	0.0023	0.00
HRP	0.0016	0.08	0.0017	0.03	0.0018	0.00	0.0018	0.00
IVP	0.0017	0.02	0.0017	0.02	0.0019	0.00	0.0020	0.00

Table 5.1: For the datasets involving monthly returns of "Industry Portfolios ($N = 10 + 1$)" and "Industry Portfolios ($N = 30$)", this table reports the out-of-sample risk for the portfolio allocation strategies described in Section 5.2 for two different training periods ($M = 60$ and $M = 120$). The column "p" shows the p-value of the ratio of each strategy's portfolio risk to Min-c benchmark's portfolio risk as it is described in Section 5.3.2.

	SPI Sectors, $N = 10$				Dow Jones, $N = 28$			
	$M = 130$		$M = 520$		$M = 130$		$M = 520$	
Strategies	Risk	p	Risk	p	Risk	p	Risk	p
MV-c	0.0035	0.00	0.0038	0.00	0.0084	0.00	0.0100	0.00
MV-c-SLCA	0.0035	0.00	0.0039	0.00	0.0085	0.00	0.0101	0.00
Min-c(*)	0.0011	*	0.0013	*	0.0015	*	0.0015	*
Min-c-SLCA	0.0011	0.46	0.0013	0.45	0.0015	0.44	0.0015	0.46
MV-c-RetHRP	0.0012	0.18	0.0014	0.25	0.0015	0.47	0.0015	0.47
MV-c-SLCA-RetHRP	0.0012	0.17	0.0014	0.28	0.0015	0.43	0.0016	0.41
EW	0.0020	0.00	0.0020	0.00	0.0024	0.00	0.0026	0.00
HRP	0.0015	0.01	0.0016	0.05	0.0019	0.07	0.0020	0.03
IVP	0.0015	0.00	0.0017	0.02	0.0019	0.04	0.0020	0.02

Table 5.2: For the datasets involving daily returns of SPI Sectors ($N = 10$) and Dow Jones stocks ($N = 28$), this table reports the out-of-sample risk for the portfolio allocation strategies described in Section 5.2 for two different training periods ($M = 130$, $M = 520$). The column "p" shows the p-value of the ratio of each strategy's portfolio risk to Min-c benchmark's portfolio risk as it is described in Section 5.3.2.

of the portfolios and is the key criterion to compare the portfolio allocation strategies' out of sample performances with the naive equally-weighted (EW) strategy in DeMiguel and Nogales (2009)'s study.

Figure 5.3 and Figure 5.4 display the out of sample Sharpe ratios of the investigated strategies as a function of the training time. Contrary to the risk performance metric comparison in Section 5.4.1, there is no clear winner or clear loser strategy here. The ranking of the portfolio allocation strategies based on their Sharpe ratios differ for different datasets and different training periods. For instance, for the dataset "$N = 10 + 1$ industry portfolios", the best performing strategy changes with different training periods. Overall, HRP and MV-c-SLCA-RetHRP can be thought as the best performing strategies although for the 7 of the 9 strategies, Sharpe ratios are quite close to each other when training period is 120 months. On the other hand, Min-c, Min-c-SLCA, MV-c-RetHRP and MV-c-SLCA-RetHRP are the top performers for the "Industry 30 portfolios" dataset. For the "SPISectors" dataset, MV-c-RetHRP and MV-c-RetHRP-SLCA are the best performing strategies for larger training periods while all the strategies but MV-c and MV-c-SLCA, are quite close to each other for shorter training periods. For three out of four datasets used, MV-c and MV-c-SLCA are clearly the worst performing strategies.

However, for the "Dow Jones" dataset, where the idiosynratic volatility are expected to be higher for individual assets compared to the other investigated datasets, MV-c and MV-c-SLCA are returning Sharpe ratios as high as the other strategies and it is the top performing portfolio for $M = 520$ training days although it is not significantly different than EW as we later see from Table 5.4.

Given there is no clear winner among the portfolio strategies in terms of out-of-sample Sharpe ratios and following DeMiguel and Nogales (2009)'s study, we take equally weighted (EW) allocation strategy as our benchmark model when we calculate the p-values for the statistical significance of the difference of the Sharpe ratios

obtained by the other strategies. The Sharpe ratios and p values are reported in Table 5.3 and Table 5.4 for training periods $M = 60$ and $M = 120$ for monthly returns and $M = 130$ and $M = 520$ for daily returns.

For the datasets involving monthly asset returns, the results are shown in Table 5.3. The dataset "Industry Portfolios, $N = 10 + 1$" is one of the datasets that DeMiguel and Nogales (2009) used to compare portfolio allocations' strategies. In Table 5.3, we replicate his experiments and our findings for the Sharpe ratio and p-values match with the results they reported in their study.

For the monthly returns dataset, the EW portfolios consistently do better than MV-c and MV-c-SLCA portfolios for both training periods. The difference is also statistically significant between both of these strategies and the EW strategy only when the training period is chosen as $M = 120$ for both datasets. The rest of the strategies, other than the strategies MV-c and MV-c-SLCA, return higher Sharpe ratios than the EW strategy for "30 Industry Portfolios" dataset. All these strategies are performing statistically better than the EW strategy when the training period is 60 months (p-values are less than 0.05). When the training period is 120, the differences between the strategies HRP, IVP and MV-c-RetHRP's Sharpe ratio and the EW's sharpe ratio are statistically significant. Also, for the same training period the EW's Sharpe ratio and the Sharpe ratios of the strategies Min-c, Min-c-SLCA and MV-c-SLCA-RetHRP are considerably different (the p values are 0.06, 0.07 and 0.06 respectively).

For the datasets involving daily asset returns, the results are given in Table 5.4. For the daily returns, the difference between EW's and the other strategies' Sharpe ratios are not dramatically significant. For the "SPI sectors" dataset, MV-c and MV-c-SLCA significantly underperforms comparing to the other strategies when the training period is 130 days. When the training period is 520 days, MV-c-RetHRP and MV-c-SLCA-RetHRP methods are seen to be the highest performing among

the others and they are considerably better than selected benchmark EW. For the "Dow Jones" dataset, EW strategy is among the top performers, however it does not significantly outperform the other strategies. Interestingly, mean-variance models MV-c and MV-c-SLCA return comparably high Sharpe ratios, especially when the training period M is 520 days. However, the difference is not statistically significant at the conventional levels, such as $p = 0.01$ or $p = 0.05$, returning $p = 0.08$.

In general, our findings resulting from the Sharpe ratio comparison align with the findings from Section 5.4.1 where we compare out-of-sample risk values. For all the datasets except for the dataset "Dow Jones", MV-c and MV-c-SLCA models yield lower Sharpe ratios. That is the effect of estimation error is so large that it erodes completely the gains from optimal diversification. Also similar to the comparison between the out-of-sample risk, we do not observe any significant improvement of using filtered covariance matrices in the optimization models such as mean-variance, minimum-variance or mean-variance with a benchmark return constraint. Similarly, the most recent portfolio allocation strategy HRP's performance is very comparable to traditional IVP's performance, and it does not consistently outperform the other allocation strategies, including EW.

On the other hand, the rest of the strategies are generally comparable to each other. Although for some of the selected datasets and some certain training periods, the selected benchmark EW significantly underperforms compared to some of the other portfolio allocation strategies, this behavior is not consistent, and there is no single allocation strategy consistently outperforming all the other allocation strategies.

Finally, using HRP allocation's portfolio return as a benchmark parameter in benchmark constrained mean-variance models (MV-c-RetHRP and MV-c-SLCA-RetHRP yield better Sharpe ratios as it yields better out of sample risk compared to the mean variance models using risk aversion parameter $\gamma = 1$ (MV-c and MV-c-SLCA).

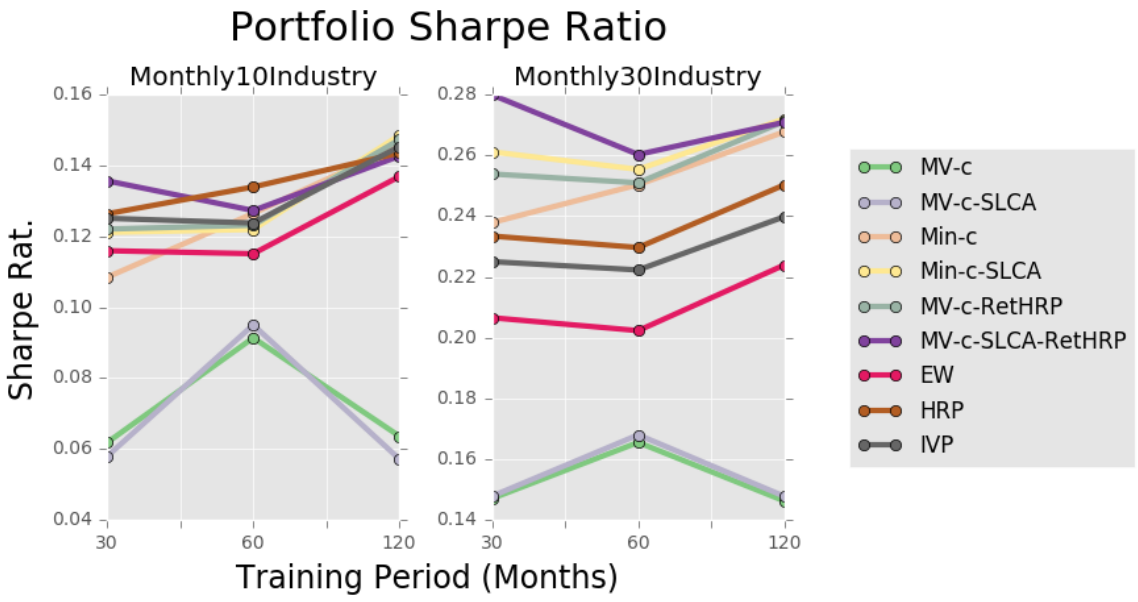


Figure 5.3: The out-of-sample Sharpe ratios for the portfolio strategies as a function of training periods for the datasets involving monthly returns of Industry Portfolios ($N = 10 + 1$) and Industry Portfolios ($N = 30$)

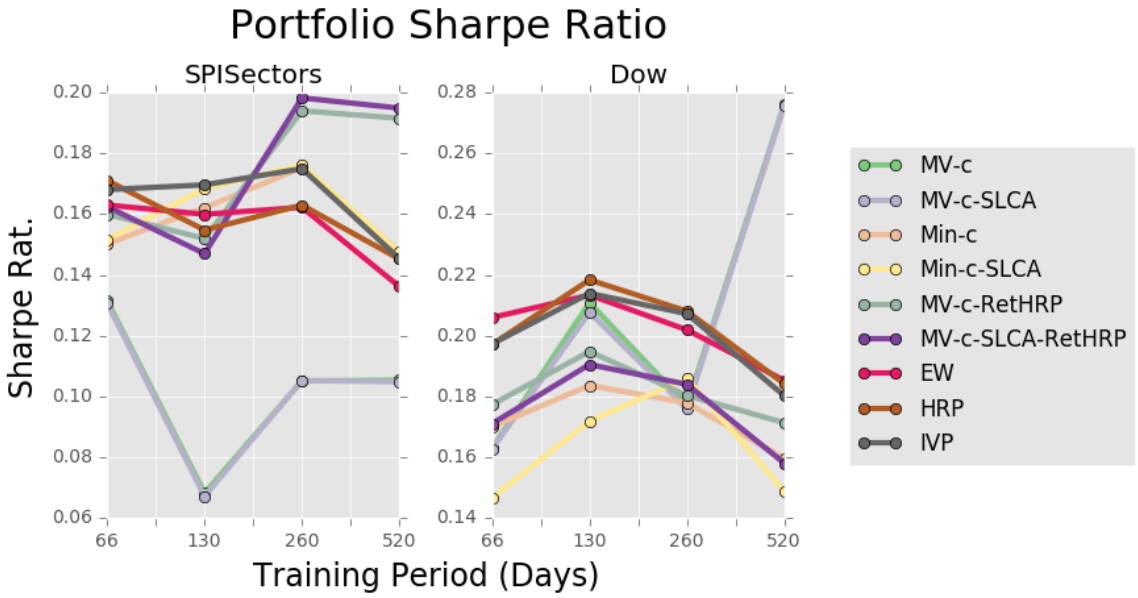


Figure 5.4: The out of sample Sharpe ratios for the portfolio strategies as a function of training periods for the datasets involving daily returns of SPI Sectors ($N = 10$) and Dow Jones stocks ($N = 28$)

	Industry Portfolios, $N = 10 + 1$				Industry Portfolios, $N = 30$			
	$M = 60$		$M = 120$		$M = 60$		$M = 120$	
Strategies	Sharpe	p	Sharpe	p	Sharpe	p	Sharpe	p
MV-c	0.0914	0.24	0.0636	0.02	0.1655	0.15	0.1462	0.01
MV-c-SLCA	0.0950	0.28	0.0571	0.02	0.1680	0.17	0.1477	0.02
Min-c	0.1264	0.33	0.1457	0.39	0.2504	0.03	0.2678	0.06
Min-c-SLCA	0.1219	0.41	0.1485	0.37	0.2554	0.04	0.2720	0.07
MV-c-RetHRP	0.1231	0.36	0.1473	0.35	0.2510	0.02	0.2712	0.04
MV-c-SLCA-RetHRP	0.1273	0.30	0.1427	0.42	0.2603	0.02	0.2709	0.06
EW(*)	0.1151	*	0.1369	*	0.2023	*	0.2239	*
HRP	0.1339	0.01	0.1436	0.20	0.2297	0.00	0.2503	0.00
IVP	0.1237	0.04	0.1453	0.05	0.2223	0.00	0.2398	0.00

Table 5.3: For the datasets involving monthly returns of "Industry Portfolios ($N = 10 + 1$)" and "Industry Portfolios ($N = 30$)", this table reports the out-of-sample Sharpe ratios for the portfolio allocation strategies described in Section 5.2. The column "p" shows the p-value of the difference between the Sharpe ratio of each strategy's from that of the EW benchmark, which is computed using the Jobson and Korkie (1981) described in Section 5.3.2.

	SPI Sectors, $N = 10$				Dow Jones, $N = 28$			
	$M = 130$		$M = 520$		$M = 130$		$M = 520$	
Strategies	Sharpe	p	Sharpe	p	Sharpe	p	Sharpe	p
MV-c	0.0682	0.04	0.1054	0.28	0.2108	0.48	0.2763	0.08
MV-c-SLCA	0.0671	0.04	0.1048	0.28	0.2077	0.47	0.2758	0.08
Min-c	0.1621	0.48	0.1476	0.39	0.1836	0.21	0.1595	0.26
Min-c-SLCA	0.1683	0.43	0.1474	0.40	0.1719	0.16	0.1487	0.20
MV-c-RetHRP	0.1520	0.41	0.1916	0.06	0.1948	0.31	0.1713	0.34
MV-c-SLCA-RetHRP	0.1470	0.36	0.1949	0.05	0.1905	0.29	0.1580	0.25
EW(*)	0.1600	*	0.1361	*	0.2134	*	0.1852	*
HRP	0.1548	0.36	0.1452	0.25	0.2184	0.37	0.1841	0.47
IVP	0.1697	0.16	0.1453	0.18	0.2139	0.48	0.1802	0.33

Table 5.4: For the datasets involving daily returns of "SPI Sectors" ($N = 10$) and "Dow Jones stocks" ($N = 28$), this table reports the out-of-sample Sharpe ratios for the portfolio allocation strategies described in Section 5.2. The column "p" shows the p-value of the difference between the Sharpe ratio of each strategy's from that of the EW benchmark, which is computed using the Jobson and Korkie (1981) described in Section 5.3.2.

5.4.3 Turnover

Turnover values can be used as a proxy estimate to the transaction costs that occur due to rebalancing of the portfolios. The higher the turnover, the more the transaction costs an investor has to pay.

Figure 5.5 and Figure 5.6 display average turnover rates of the investigated strategies as a function of the training period. For both type of datasets, the datasets involving monthly return and the datasets involving the daily return, the turnover rates decrease as the training period increases for all the strategies but for the EW strategy. This can be interpreted as a sign of an increase in portfolio stability with the time. The EW strategy's turnover rate is not affected by the increase in training periods as it does not estimate any parameter to find the portfolio weights. The highest turnover rates belong to MV-c and MV-c-SLCA models while the lowest turnover rates are observed for the EW and IVP strategies. The IVP strategy returns the lowest turnover rate as the training period increases. Table 5.5 and Table 5.6 report some of the selected training periods for the datasets. All of the p-values of the difference between the average turnover values of each strategy's from that of the EW benchmark show that the differences between the turnover rates are significant.

5.5 Conclusion

In this chapter, we compared the performance of 9 models for the portfolio allocation, including the models with the filtered covariance matrices using hierarchical clustering and HRP portfolio allocation strategy. The comparison is conducted by using four different datasets in a widely accepted experimental design setting. We show that using filtered covariance matrix in short sale constrained models does not create a significant performance change in the models. Our second finding is the recently developed portfolio allocation strategy "HRP" performs very closely to the

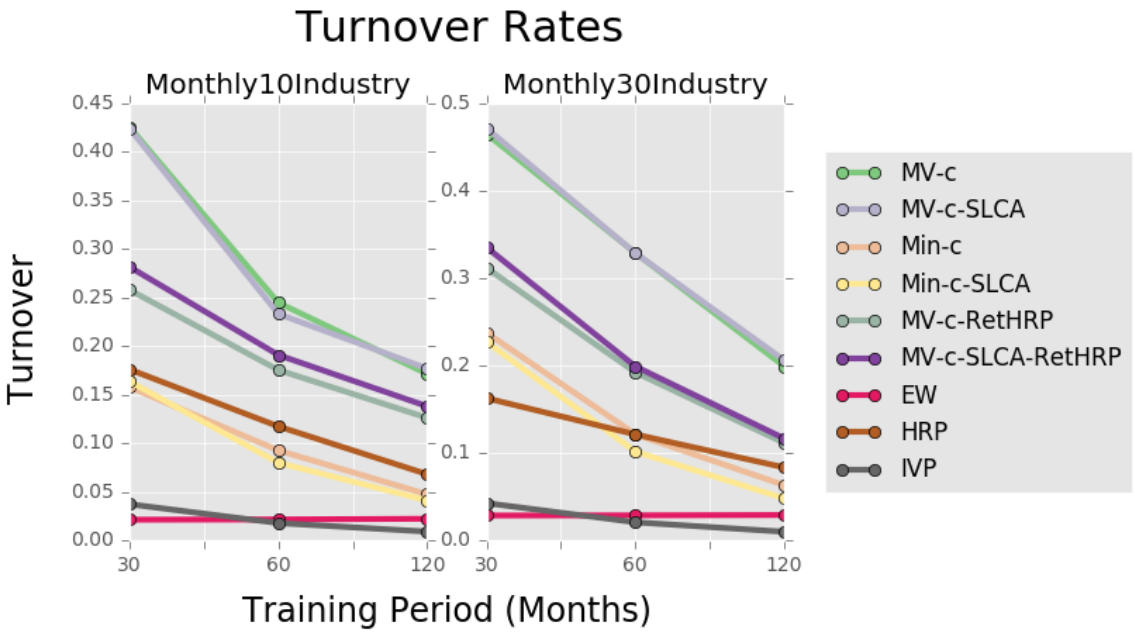


Figure 5.5: The average turnover rates for the portfolio strategies as a function of training periods for the datasets involving monthly returns of Industry Portfolios ($N = 10 + 1$) and Industry Portfolios ($N = 30$)

	Industry Portfolios, $N = 10 + 1$				Industry Portfolios, $N = 30$			
	$M = 60$		$M = 120$		$M = 60$		$M = 120$	
Strategies	Turnover	p	Turnover	p	Turnover	p	Turnover	p
MV-c	0.2447	0.00	0.1713	0.00	0.3288	0.00	0.1991	0.00
MV-c-SLCA	0.2331	0.00	0.1769	0.00	0.3286	0.00	0.2060	0.00
Min-c	0.0925	0.00	0.0473	0.00	0.1212	0.00	0.0630	0.00
Min-c-SLCA	0.0798	0.00	0.0409	0.00	0.1014	0.00	0.0478	0.00
MV-c-RetHRP	0.1753	0.00	0.1261	0.00	0.1921	0.00	0.1113	0.00
MV-c-SLCA-RetHRP	0.1908	0.00	0.1379	0.00	0.1989	0.00	0.1164	0.00
EW	0.0214	*	0.0221	*	0.0284	*	0.0288	*
HRP	0.1172	0.00	0.0680	0.00	0.1211	0.00	0.0833	0.00
IVP	0.0176	0.00	0.0089	0.00	0.0205	0.00	0.0096	0.00

Table 5.5: For the datasets involving monthly returns of "Industry Portfolios ($N = 10 + 1$)" and "Industry Portfolios ($N = 30$)", this table reports the average turnover rates for the portfolio allocation strategies described in Section 5.2. The column "p" shows the p-value of the difference between the average turnover values of each strategy's from that of the EW benchmark, which is computed using the t-tests.

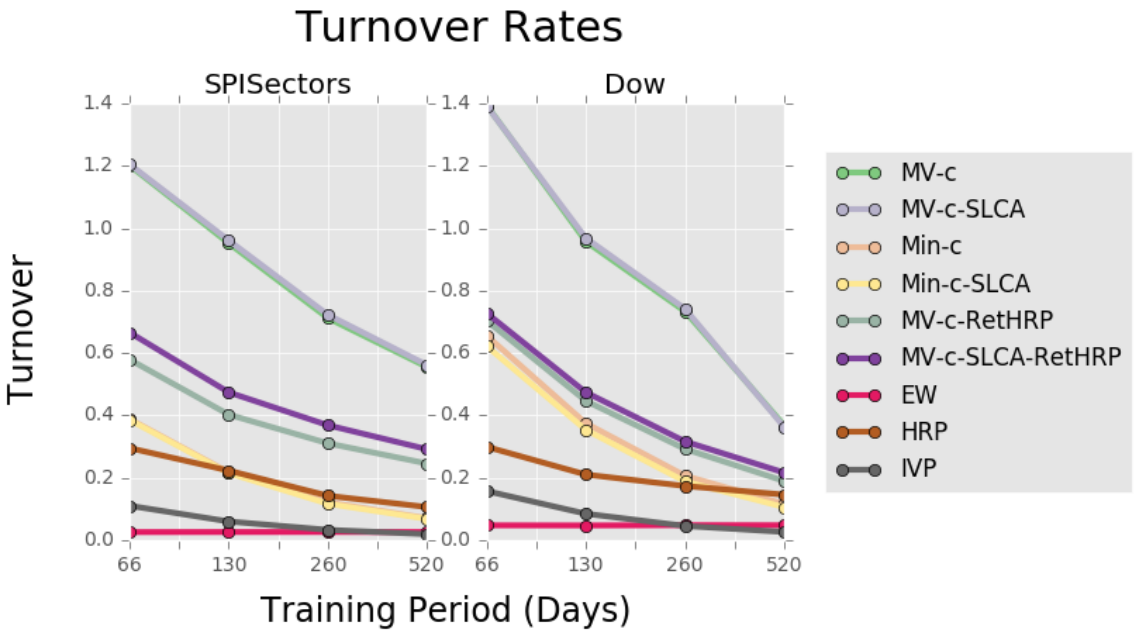


Figure 5.6: The average turnover rates for the portfolio strategies as a function of training periods for the datasets involving daily returns of SPI Sectors ($N = 10$) and Dow Jones stocks ($N = 28$)

	SPI Sectors, $N = 10$				Dow Jones, $N = 28$			
	$M = 130$		$M = 520$		$M = 130$		$M = 520$	
Strategies	Turnover	p	Turnover	p	Turnover	p	Turnover	p
MV-c	0.9518	0.00	0.5550	0.00	0.9576	0.00	0.3680	0.00
MV-c-SLCA	0.9610	0.00	0.5614	0.00	0.9686	0.00	0.3628	0.00
Min-c	0.2172	0.00	0.0702	0.00	0.3748	0.00	0.1156	0.00
Min-c-SLCA	0.2148	0.00	0.0682	0.00	0.3537	0.00	0.1058	0.00
MV-c-RetHRP	0.4025	0.00	0.2456	0.00	0.4465	0.00	0.1895	0.00
MV-c-SLCA-RetHRP	0.4739	0.00	0.2915	0.00	0.4738	0.00	0.2165	0.00
EW	0.0261	*	0.0261	*	0.0472	*	0.0478	*
HRP	0.2231	0.00	0.1062	0.00	0.2103	0.00	0.1459	0.00
IVP	0.0600	0.00	0.0194	0.00	0.0847	0.00	0.0263	0.00

Table 5.6: For the datasets involving daily returns of "SPI Sectors ($N = 10$)" and "Dow Jones stocks ($N = 28$)", this table reports the average turnover rates for the portfolio allocation strategies described in Section 5.2. The column "p" shows the p-value of the difference between the average turnover values of each strategy's from that of the EW benchmark, which is computed using the t-tests.

traditional risk parity allocation strategy, and it does not consistently outperform the classical strategies such as the equally-weighted approach, mean-variance models and minimum-variance models. Also, we verified the general poor out-of-sample performance of the mean-variance models, which are in general the worst performers because of the estimation errors. In summary, there is no single model that consistently delivers an out-of-sample risk or a Sharpe ratio or a turnover rate consistently outperforming the other strategies. The performance of the methods are generally affected by the training period of the models, the frequency of the returns and the constraints in the models. The estimation error free equally-weighted approach is still an attractive choice for investors with its low turnover rate and competitive out-of-sample performance.

Bibliography

Mansoor Hamood Al-Harthy. Stochastic oil price models: comparison and impact. *The Engineering Economist*, 52(3):269–284, 2007.

A. Alexander, T. F. Coleman, and Y. Lo. Minimizing Var and CVaR for a portfolio of derivatives. *Journal of Banking Finance*, 30(2):583–605, 2006.

A Almansoori and N Shah. Design and operation of a stochastic hydrogen supply chain network under demand uncertainty. *International journal of hydrogen energy*, 37(5): 3965–3977, 2012.

Ali Almansoori and N Shah. Design and operation of a future hydrogen supply chain: snapshot model. *Chemical Engineering Research and Design*, 84(6):423–438, 2006.

P. Artzner, F. Delbaen, J. M. Eber, and D. Heath. Coherent measures of risk. *Math. Finance*, 9:203–228, 1999.

Onur Babat, Ali Esmaili, Joshua D Isom, Camilo Mancilla, and Luis F Zuluaga. Systematic prioritization of sensor improvements in an industrial gas supply network. *International Journal of Chemical Engineering*, 2017, 2017a.

Onur Babat, Juan Vera, and Luis Zuluaga. Computing near-optimal value-at-risk portfolios using integer programming techniques. Technical report, Lehigh University, 2017b. Under review for European Journal of Operations Research.

Vijay S Bawa, Stephen J Brown, and Roger W Klein. Estimation risk and optimal portfolio choice. *NORTH-HOLLAND PUBL. CO., N. Y., 190 pp*, 1979.

S. Bazak and A. Shapiro. Value-at-Risk based risk management: Optimal policies and asset prices. *Review of Financial Studies*, 14(2):371–405, 2001.

Fabio Bellini and Valeria Bignozzi. Elicitable risk measures. *Available at SSRN 2334746*, 2013.

Stefano Benati and Romeo Rizzi. A mixed integer linear programming formulation of the optimal mean/value-at-risk portfolio problem. *European Journal of Operational Research*, 176(1):423–434, 2007.

Dimitris Bertsimas and Romy Shioda. Algorithm for cardinality-constrained quadratic optimization. *Computational Optimization and Applications*, 43(1):1–22, 2009.

Daniel Bienstock. Computational study of a family of mixed-integer quadratic programming problems. *Mathematical Programming*, 74(2):121–140, 1996.

J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer Series in Operations Research and Financial Engineering. Springer, 2nd edition, 2011.

Robert E. Bixby. Solving real-world linear programs: A decade and more of progress. *Operations Research*, 50(1):3–15, 2002.

F. Black and R. Litterman. Global portfolio optimization. *Financial Analysts Journal*, 48 (5):28–43, 2001.

V. Boasson, E. Boasson, and Z. Zhou. Portfolio optimization in a mean-semivariance framework. *Investment management and financial innovations*, 8(3):58–68, 2011.

Z. Bodie, A. Kane, and A. J. Markus. *Investments*. McGraw-Hill, 2005.

B. Borchers and J. E. Mitchell. An improved branch and bound algorithm for mixed integer nonlinear programs. *Computers and Operations Research*, 21(4):359–367, 1994.

Emanuele Borgonovo and Elmar Plischke. Sensitivity analysis: a review of recent advances. *European Journal of Operational Research*, 2015.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Maria Debora Braga. The traditional approach to asset allocation. In *Risk-Based Approaches to Asset Allocation*, pages 3–15. Springer, 2016.

Johan Bring. How to standardize regression coefficients. *The American Statistician*, 48(3): 209–213, 1994.

Bruce L Brown and Suzanne B Hendrix. Partial correlation coefficients. *Wiley StatsRef: Statistics Reference Online*, 2005.

Dan G Cacuci, Mihaela Ionescu-Bujor, and Ionel Michael Navon. *Sensitivity and uncertainty analysis, volume II: applications to large-scale systems*, volume 2. CRC Press, 2005.

Andrew JG Cairns. A discussion of parameter and model uncertainty in insurance. *Insurance: Mathematics and Economics*, 27(3):313–330, 2000.

M. C. Campi and G. Calafiore. Uncertain convex programs: Randomized solutions and confidence levels. *Math. Programming*, 102(1):25–46, 2005.

E. Cetinkaya and A. Thiele. Data-driven portfolio management with quantile constraints. Technical report, Lehigh University, 2014. Available at `http://engineered.typepad.com/CetinkayaThiele-Data-Driven-Portfolio-Management-Quantile-Constraints.pdf`.

Vijay K Chopra, William T Ziemba, et al. The effect of errors in means, variances, and covariances on optimal portfolio choice. *The Kelly Capital Growth Investment Criterion: Theory and Practice*, 3:249, 2011.

Thomas Coleman, Mary Ann Branch, and Andrew Grace. *Optimization Toolbox for Use with MATLAB: User's Guide, Version 2*. Math Works, Incorporated, 1999.

Thomas Conlon, Heather J Ruskin, and Martin Crane. Random matrix theory and fund of funds portfolio optimisation. *Physica A: Statistical Mechanics and its applications*, 382(2):565–576, 2007.

Rama Cont, Romain Deguest, and Giacomo Scandolo. Robustness and sensitivity analysis of risk measurement procedures. *Quantitative Finance*, 10(6):593–606, 2010.

American Chemistry Council. Industrial gases energy and environmental benefits. `http://www.americanchemistry.com/ProductsTechnology/Industrial-Gases/Industrial-Gases-Energy-and-Environmental-Benefits.pdf`, 2012.

J. Daníelsson. *Financial Risk Forecasting*. Wiley, 2011.

G. Darbha. Value-at-Risk for fixed income portfolios. Technical report, National Stock Exchange, Mumbai, India, 2001. http://www.igidr.ac.in/~{}susant/FIXEDINCOME/Darbha2001_var.pdf.

D. P. de Farias and B. Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. *Math. Oper. Res.*, 29(3):462–478, 2004.

Victor DeMiguel and Francisco J Nogales. Portfolio selection with robust estimation. *Operations Research*, 57(3):560–577, 2009.

Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial Studies*, 22(5): 1915–1953, 2009.

Ye Ding, Xiao Feng, and Khim H Chu. Optimization of hydrogen distribution systems with pressure constraints. *Journal of Cleaner Production*, 19(2):204–211, 2011.

A. K. Dixit and R. S. Pindyck. *Investment Under Uncertainties*. Princeton University Press, 1994.

Ricardo Dunia, S Joe Qin, Thomas F Edgar, and Thomas J McAvoy. Identification of faulty sensors using principal component analysis. *AIChE Journal*, 42(10):2797–2812, 1996.

L. El Ghaoui, M. Oks, and F. Oustry. Worst-case Value-at-Risk and robust portfolio optimization: A conic programming approach. *Operations Research*, 51(4):543–556, 2003.

Edwin J Elton, Martin J Gruber, Stephen J Brown, and William N Goetzmann. *Modern portfolio theory and investment analysis*. John Wiley & Sons, 2009.

E. Erdogan and G. Iyengar. Ambiguous chance constrained problems and robust optimization. *Math. Programming Ser. B*, 107(1–2):37–61, 2006.

EY. Us oil and gas reserves study. Technical report, Ernst & Young, 2015. URL http://www.ey.com/Publication/vwLUAssets/EY-us-oil-and-gasreserves-study-2015/$FILE/EY-us-oil-and-gas-reserves-study-2015.pdf.

S. Fang and S. Puthenpura. *Linear Optimization and Extensions: Theory and Algorithms*. Prentice Hall, 1993.

M. Feng, A. Wächter, and J. Staum. Practical algorithms for Value-at-Risk portfolio optimization problems. *Quantitative Finance Letters*, 3(1):1–9, 2015. Available at `http://www.tandfonline.com/doi/pdf/10.1080/21649502.2014.995214`.

Anthony V Fiacco. Sensitivity analysis for nonlinear programming using penalty methods. *Mathematical programming*, 10(1):287–311, 1976.

O. E. Flippo and A. H. G. Rinnooy-Kan. Decomposition in general mathematical programming. *Mathematical Programming*, 60:361–382, 1993.

André Fonseca, Vítor Sá, Hugo Bento, Manuel LC Tavares, Gilberto Pinto, and Luísa ACN Gomes. Hydrogen distribution network optimization: a refinery case study. *Journal of Cleaner Production*, 16(16):1755–1763, 2008.

Ken French. Us research returns, 2017. URL `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html#Research`.

R. M. Freund. Benders decomposition methods for structured optimization, including stochastic optimization. Technical report, Massachusetts Institute of Technology, 2004. available at `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.3003&rep=rep1&type=pdf`.

A. A. Gaivoronski and G. Pflug. Value-at-Risk in portfolio optimization: Properties and computational approach. *Journal of Risk*, 7(2):1–31, 2004.

Paul Geladi and Bruce R Kowalski. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17, 1986.

P. Glasserman, P. Heidelberger, and P. Shahabuddin. Variance reduction techniques for estimating Value-at-Risk. *Management Science*, 46(10):1349–1364, 2000.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

Donald Goldfarb and Garud Iyengar. Robust portfolio selection problems. *Mathematics of operations research*, 28(1):1–38, 2003.

John C Gower and GJS Ross. Minimum spanning trees and single linkage cluster analysis. *Applied statistics*, pages 54–64, 1969.

José Luis Guiñón, Emma Ortega, José García-Antón, and Valentín Pérez-Herranz. Moving average and savitzki-golay smoothing filters using mathcad. *Papers ICEE*, 2007, 2007.

Lijie Guo and Jianxin Kang. A hybrid process monitoring and fault diagnosis approach for chemical plants. *International Journal of Chemical Engineering*, 2015, 2015.

Dion Harmon, Blake Stacey, Yavni Bar-Yam, and Yaneer Bar-Yam. Networks of economic market interdependence and systemic risk. *arXiv preprint arXiv:1011.3707*, 2010.

C. C. Heyde, S. G. Kou, and X. H. Peng. What is a good risk measure: Bridging the gaps between data, coherent risk measures, and insurance risk measures. Technical report, Columbia University, 2006. `http://192.147.69.84/bank/analytical/cfr/2007/apr/KOU_hkpv1.pdf`.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Toshimitsu Homma and Andrea Saltelli. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):1–17, 1996.

Ronald A. Howard. Uncertainty about probability: A decision analysis perspective. *Risk Analysis*, 8(1):91–98, 1988. ISSN 1539-6924. doi: 10.1111/j.1539-6924.1988.tb01156.x. URL `http://dx.doi.org/10.1111/j.1539-6924.1988.tb01156.x`.

P. J. Huber and E. M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley, 2009.

Joshua D Isom and Robert E LaBarre. Process fault detection, isolation, and reconstruction by principal component pursuit. In *2011 American Control Conference*, 2011.

Ravi Jagannathan and Tongshu Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, 58(4):1651–1684, 2003.

Robert Jarrow and Feng Zhao. Downside loss aversion and portfolio management. *Management Science*, 52(4):558–566, 2006.

Yunqiang Jiao, Hongye Su, Weifeng Hou, and Zuwei Liao. Optimization of refinery hydrogen network based on chance constrained programming. *Chemical Engineering Research and Design*, 90(10):1553–1567, 2012.

Yunqiang Jiao, Hongye Su, Weifeng Hou, and Pu Li. Design and optimization of flexible hydrogen systems in refineries. *Industrial & Engineering Chemistry Research*, 52(11): 4113–4131, 2013.

J Dave Jobson and Bob M Korkie. Performance hypothesis testing with the sharpe and treynor measures. *The Journal of Finance*, 36(4):889–908, 1981.

JD Jobson, Bob Korkie, and Vinod Ratti. Improved estimation for markowitz portfolios using james-stein type estimators. In *Proceedings of the American Statistical Association, Business and Economics Statistics Section*, volume 41, pages 279–284. American Statistical Association Washington DC, 1979.

P. Jorion. *Value at Risk: The New Benchmark for managing financial risk.* McGraw–Hill, New York, 2001.

N. H. Josephy and A. D. Aczel. A statistically optimal estimator of semivariance. *European Journal of Operations Research*, 67(2):267–271, 1993.

G. Kaplanski and Y. Koll. Var risk measures vs traditional risk measures: An analysis and survey. *Journal of Risk*, 4(3):43–68, 2002.

Jiyong Kim, Younghee Lee, and Il Moon. Optimization of a hydrogen supply chain under demand uncertainty. *International Journal of Hydrogen Energy*, 33(18):4715–4729, 2008.

SJ Kline. The purposes of uncertainty analysis. *Journal of Fluids Engineering*, 107(2): 153–160, 1985.

Rainer Kolisch and Arno Sprecher. PSPLIB-a project scheduling problem library: OR software–ORSEP operations research software exchange program. *European Journal of Operational Research*, 96(1):205–216, 1997.

H. Konno and H. Yamazaki. Mean-absolute deviation portfolio optimization model and its applications to Tokyo Stock Market. *Management Science*, 37(5):519–531, 1991.

Pavlo Krokhmal, Jonas Palmquist, and Stanislav Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of Risk*, 4:11–27, 2002.

Harold W Kuhn. Nonlinear programming: a historical view. In *Traces and Emergence of Nonlinear Programming*, pages 393–414. Springer, 2014.

A Kumar, G Gautami, and S Khanam. Hydrogen distribution in the refinery using mathematical modeling. *Energy*, 35(9):3763–3772, 2010.

Andrew Kusiak and Zhe Song. Sensor fault detection in power plants. *Journal of Energy Engineering*, 135(4):127–137, 2009.

N. Larsen, H. Mausser, and S. Uryasev. Algorithms for optimization of Value-at-Risk. In P. Pardalos and V. K. Tsitsiringos, editors, *Financial Engineering, e-Commerce and Supply Chain*, page 129?157. Kluwer Academic Publishers, 2002.

A. M. Law and W. D. Kelton. *Simulation modeling and analysis*. McGraw-Hill, 2000.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Changkyu Lee, Sang Wook Choi, and In-Beum Lee. Sensor fault identification based on time-lagged pca in dynamic processes. *Chemometrics and Intelligent Laboratory Systems*, 70(2):165–178, 2004.

Miguel A. Lejeune. Portfolio optimization with combinatorial and downside return constraints. In Luis F. Zuluaga and Tams Terlaky, editors, *Modeling and Optimization: Theory and Applications*, volume 62 of *Springer Proceedings in Mathematics & Statistics*, pages 31–50. Springer New York, 2013. ISBN 978-1-4614-8986-3. doi: 10.1007/978-1-4614-8987-0_2. URL `http://dx.doi.org/10.1007/978-1-4614-8987-0_2`.

Deyi Li. Artificial intelligence with uncertainty. In *Computer and Information Technology, 2004. CIT'04. The Fourth International Conference on*, pages 2–2. IEEE, 2004.

Chang-Chun Lin. Comments on a mixed integer linear programming formulation of the optimal mean/value-at-risk portfolio problem. *European Journal of Operational Research*, 194(1):339–341, 2009.

M. S. Lobo. *Robust and convex optimization with applications to finance*. PhD thesis, Department of Electrical Engineering, Stanford University, 2000.

Marcos López de Prado. Building diversified portfolios that outperform out of sample. *The Journal of Portfolio Management*, 42(4):59–69, 2016.

Junyi Lou, Zuwei Liao, Binbo Jiang, Jingdai Wang, and Yongrong Yang. Robust optimization of hydrogen network. *International Journal of Hydrogen Energy*, 39(3):1210–1219, 2014.

Leonardo Lozano and Andrs L. Medaglia. On an exact method for the constrained shortest path problem. *Computers & Operations Research*, 40(1):378 – 384, 2013. ISSN 0305-0548. doi: http://dx.doi.org/10.1016/j.cor.2012.07.008. URL `http://www.sciencedirect.com/science/article/pii/S0305054812001530`.

Renata Mansini, Wlodzimierz Ogryczak, and M. Grazia Speranza. Twenty years of linear programming based portfolio optimization. *European Journal of Operational Research*, 2013. ISSN 0377-2217. doi: http://dx.doi.org/10.1016/j.ejor.2013.08.035. URL `http://www.sciencedirect.com/science/article/pii/S0377221713007194`.

Rosario N Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11(1):193–197, 1999.

H. Markowitz, P. Todd, G. Xu, and Y. Yamane. Computation of mean-semivariance efficient sets by the critical line algorithm. *Annals of Operations Research*, 45(1):307–317, 1993.

Harry Markowitz. Portfolio selection*. *The journal of finance*, 7(1):77–91, 1952.

Gautier Marti, Frank Nielsen, Mikołaj Bińkowski, and Philippe Donnat. A review of two decades of correlations, hierarchies, networks and clustering in financial markets. *arXiv preprint arXiv:1703.00485*, 2017.

Nasir Mehranbod, Masoud Soroush, and Chanin Panjapornpon. A method of sensor fault detection and identification. *Journal of Process Control*, 15(3):321–339, 2005.

A. Mehrotra and M. A. Trick. A branch-and-price approach for graph multi-coloring. *Extending the Horizons: Advances in Computing, Optimization, and Decision Technologies Operations Research/Computer Science Interfaces Series*, 37:15–29, 2007.

Don Merritt, Andre de San Miguel, et al. Portfolio optimization using efficient frontier theory. In *SPE Asia Pacific Conference on Integrated Modelling for Asset Management*. Society of Petroleum Engineers, 2000.

A. Meucci. *Risk and asset allocation*. Springer, 2007.

R. O. Michaud. *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation*. Oxford University Press, 1998.

Richard O Michaud. The markowitz optimization enigma: Is optimized optimal? *ICFA Continuing Education Series*, 1989(4):43–54, 1989.

K. Natarajan, M. Sim, and D. Pachamanova. Constructing risk measures from uncertainty sets. *Operations Research*, 57(5), 2009.

G. L. Nemhauser and L. A. Wolsey. *Integer and combinatorial optimization*. Wiley, 1988.

Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

Wyoming Oil and Gas Conservation Commission. Wyoming oil and gas conservation commission state of wyoming approved permits to drill. Technical report, 2015. URL `http://wogcc.state.wy.us/DumpApds.cfm`.

MM Orman, TE Duggan, et al. Applying modern portfolio theory to upstream investment decision making. *Journal of petroleum technology*, 51(03):50–53, 1999.

Ester Pantaleo, Michele Tumminello, Fabrizio Lillo, and Rosario N Mantegna. When do improved covariance matrix estimators enhance portfolio optimization? an empirical comparative study of nine estimators. *Quantitative Finance*, 11(7):1067–1080, 2011.

António Quintino, João Carlos Lourenço, and Margarida Catalão-Lopes. An integrated risk management model for an oil and gas company. *Business Administration*, page 144, 2013.

Thomas Raffinot. Hierarchical clustering based asset allocation. *Browser Download This Paper*, 2016.

R. T. Rockafellar and S. Uryasev. Optimization of Conditional Value-at-Risk. *Journal of Risk*, 2:21–41, 2000.

T. R. Rockafellar, S. P. Uryasev, and M. Zabarankin. Generalized deviations in risk analysis. Technical Report 2004-4, University of Florida, 2004.

Andrzej Ruszczyński. Decomposition methods in stochastic programming. *Mathematical programming*, 79(1-3):333–353, 1997.

Andrea Saltelli, Stefano Tarantola, Francesca Campolongo, and Marco Ratto. *Sensitivity analysis in practice: a guide to assessing scientific models*. John Wiley & Sons, 2004.

Andrea Saltelli, Marco Ratto, Stefano Tarantola, and Francesca Campolongo. Sensitivity analysis for chemical models. *Chemical reviews*, 105(7):2811–2828, 2005.

Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.

Sergey Sarykalin, Gaia Serraino, and Stan Uryasev. Value-at-risk vs. conditional value-at-risk in risk management and optimization. *Tutorials in Operations Research. INFORMS, Hanover, MD*, 2008.

Ronald W Schafer. What is a savitzky-golay filter?[lecture notes]. *Signal Processing Magazine, IEEE*, 28(4):111–117, 2011.

Jorge A Sefair, Carlos Y Méndez, Onur Babat, Andrés L Medaglia, and Luis F Zuluaga. Linear solution schemes for mean-semivariance project portfolio selection problems: An application in the oil and gas industry. *Omega*, 68:39–48, 2017.

P Seferlis and AN Hrymak. Sensitivity analysis for chemical process optimization. *Computers & chemical engineering*, 20(10):1177–1200, 1996.

George Lennox Sharman Shackle. *Uncertainty in Economics and other Reflections*. CUP Archive, 1955.

Enrique Sira. Semivariance as real project portfolio optimisation criteria–an oil and gas industry application. *International journal of global energy issues*, 26(1):43–61, 2006.

Ilya M Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1):271–280, 2001.

SB Suslick and DJ Schiozer. Risk analysis applied to petroleum exploration and production: an overview. *Journal of Petroleum Science and Engineering*, 44(1):1–9, 2004.

Vincenzo Tola, Fabrizio Lillo, Mauro Gallegati, and Rosario N Mantegna. Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, 32(1):235–258, 2008.

Michele Tumminello, Fabrizio Lillo, and Rosario N Mantegna. Correlation, hierarchies, and networks in financial markets. *Journal of Economic Behavior & Organization*, 75(1): 40–58, 2010.

PA Tyler, JR McVean, et al. Significance of project risking methods on portfolio optimization models. *Society of Petroleum Engineers Inc*, 2001.

Susara A van den Heever and Ignacio E Grossmann. A strategy for the integration of production planning and reactive scheduling in the optimization of a hydrogen supply network. *Computers & Chemical Engineering*, 27(12):1813–1839, 2003.

Venkat Venkatasubramanian, Raghunathan Rengaswamy, Kewen Yin, and Surya N Kavuri. A review of process fault detection and diagnosis: Part i: Quantitative model-based methods. *Computers & chemical engineering*, 27(3):293–311, 2003.

A. Verma and T. Coleman. Global optimization of VaR (Value-at-Risk) portfolios using continuation methods. In *Workshop on Optimization in Finance*. Center for International Mathematics, School of Economics, University of Coimbra, Portugal, 2005.

Harvey M Wagner. Global sensitivity analysis. *Operations Research*, 43(6):948–969, 1995.

Michael R Walls. Combining decision analysis and portfolio management to improve project selection in the exploration and production firm. *Journal of Petroleum Science and Engineering*, 44(1):55–65, 2004.

J. Wang and W. L. Hwang. A fuzzy set approach for R&D portfolio selection using a real options valuation model. *Omega*, 35(3):247–257, 2007.

M. Woodside-Oriakhi, C. Lucas, and J.E. Beasley. Portfolio rebalancing with an investment

horizon and transaction costs. *Omega*, 41(2):406 – 420, 2013. ISSN 0305-0483. URL http://www.sciencedirect.com/science/article/pii/S0305048312000680.

David Wozabal. Value-at-risk optimization using the difference of convex algorithm. *OR spectrum*, 34(4):861–883, 2012.

David Wozabal, Ronald Hochreiter, and Georg Ch Pflug. A difference of convex formulation of value-at-risk constrained optimization. *Optimization*, 59(3):377–400, 2010.

JIAO Yunqiang, SU Hongye, LIAO Zuwei, and HOU Weifeng. Modeling and multi-objective optimization of refinery hydrogen network. *Chinese Journal of Chemical Engineering*, 19(6):990–998, 2011.

Appendix A

Extra set of constraints and tables for Chapter 2

A.1 Precedence constraints for the oil and gas case study

$$\text{Precedence Relations} = \left\{ x \in \{0,1\}^n : \begin{array}{r} x_5 \leq x_{25}; \\ x_8 \leq x_{24}; \\ x_5 \leq x_{24}; \\ x_8 \leq x_{21}; \\ x_4 \leq x_{20}; \\ x_2 \leq x_{19}; \\ x_2 \leq x_{17}; \\ x_1 \leq x_{16}; \end{array} \right\} \tag{A.1}$$

A.2 The precedence relations and average net-present values (NPV) of the projects

Projects	**Avg. NPV**		
	$m = 1000$	$m = 500$	$m = 100$
1	305.897	300.329	293.160
2	460.993	454.685	446.099
3	0.107	0.103	0.098
4	67.265	65.698	64.094
5	97.901	95.899	93.731
6	0.036	0.034	0.028
7	125.989	123.594	120.963
8	106.834	104.655	102.824
9	6.128	5.979	5.766
10	8.877	8.708	8.156
11	30.411	29.568	28.955
12	159.684	156.401	153.327
13	6.681	6.562	6.471
14	0.057	0.049	0.040
15	101.636	100.084	98.133
16	37.272	36.460	35.702
17	33.162	32.495	31.192
18	18.138	17.841	17.388
19	66.076	64.928	63.245
20	10.962	10.734	10.428
21	19.332	18.883	18.413
22	17.926	17.435	16.924
23	18.780	18.284	17.795
24	20.957	20.491	19.857
25	330.199	330.508	314.499
26	1.693	1.662	1.628
27	5.574	5.551	5.258

Table A.1: Average project NPV for different sample sizes

Biography

Onur Babat was born in Erzurum, Turkey in 1988 to Seval and Cemal Babat. In 2011, he received his B.Sc. degree in Industrial Engineering with the highest honors from TOBB University of Economics and Technology, Ankara, Turkey. Onur has joined the Industrial and Systems Engineering Department at Lehigh University in the same year to pursue his doctoral degree. In 2014, he obtained his M.Eng. degree in Management Science and Engineering from the same department. He was a recipient of the Rossin Doctoral Fellowship at Lehigh in 2015. He served as the president for the INFORMS Student Chapter and the Turkish Student Club at Lehigh University in 2014. Onur will be joining the Quantitative Advisory Services practice at Ernst & Young in New York, NY as a Senior in Summer 2017.