Robust Fragmentation: A Data-Driven Approach to Decision-Making under Distributional Ambiguity

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Jeffrey Moulton

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Professor Gilad Lerman
Professor Shuzhong Zhang

July, 2016

Acknowledgements

I would like to thank my co-advisor, Shuzhong Zhang, for his support and encouragement over the past several years. His wealth of knowledge and willingness to share it were crucial to my success. I would also like to thank my advisor, Gilad Lerman, for his advice and support. I appreciate his intentions to help his students achieve the best and his flexibility and determination in getting them there. I want to thank all of the other faculty and students who played a role in my education and research. Specifically, I am grateful to Xiaobo Li for his ideas and teamwork. I am also grateful to faculty Daniel Spirn, Adam Rothman, Fadil Santosa, and Yuhong Yang for serving on committees, reviewing, and providing helpful feedback and mentoring.

I also want to thank the University of Minnesota Informatics Institute and the Mathematics Department for helping fund my research. I appreciate all of the friends who have provided a support network in which I could thrive. I am especially grateful to Brittany Baker and Madeline Schrier for their friendship and advice in challenging times. Finally, I would like to thank my parents for their unconditional love and support, and their willingness to help me get through obstacles.

Dedication

I wish to dedicate this work to the memory of my grandfather, Eugene Moulton, who would have been very proud to have the academic tradition continued.

Abstract

Decision makers often must consider many different possible future scenarios when they make a decision. A manager must choose inventory levels to maximize profit when the demand is unknown, investors choose a portfolio to maximize gain in the face of uncertain stock price fluctuations, and service providers must pick facility locations to minimize distances traveled where the location of the next customer is random. When uncertainty is involved, the desired outcome will be affected by factors and variables outside the decider's control and knowledge at the time of the decision. Such factors of uncertainty are hard to account for, especially when exact information about the uncertainty is unknown. It may be difficult to account for all possibilities or to estimate how likely each scenario is. There may be an infinite number of scenarios, too many for the human brain to contemplate. Generally, a precise distribution of these factors is unavailable. The unknown data, which can be quantified as a vector of parameters, follows an unknown distribution leading to the term distributional ambiguity. The decision maker may only have a sample of past events to guide them in balancing the costs and benefits of every possibility. Such limited historical data can help to provide guidance, but how best to use it? One would desire a decision strategy that will efficiently use the data and perform well in terms of the expectation, but also be robust to possible noise, changing conditions, and small sample size. To this end, we develop robust fragmentation (RF), a data-driven approach to decision-making under distributional ambiguity.

We consider a stochastic programming framework, where the decision maker aims to optimize an expected outcome. However, we assume that the governing distribution of the random parameters affecting the outcome is unknown. Only a historical sample of past realizations of such parameters is available. Our goal is to leverage the historical data to effectively and robustly approximate the true problem. This is done by fragmenting the support of the data into pieces to dissect the structure. We reduce the data by replacing it with summary statistics by region. These parameters are used to construct an ambiguity set for the distribution. The ambiguity set consists of all distributions which would satisfy the same regional reduction. We therefore reduce the

problem size and avoid overfitting to the training sample. Our approach allows for two types of ambiguity: ambiguity in support and ambiguity in probability. After constructing the ambiguity set, we consider the worst case expectation to provide distributional robustness. Constraining the distribution regionally allows us to capture detailed information about the distribution and keeps the approach from being too conservative. The ambiguity may be tailored to the structure, amount, and reliability of the data.

Robust fragmentation is a generalization and extension of several classical and newer approaches to approximating a stochastic program, including the sample average approximation (SAA), moments-based distributionally robust optimization (MDRO), and distributionally robust optimization based on probabilities of individual events. We demonstrate how RF conceptually fits into the greater picture and provides an adaptive way of balancing the benefits and drawbacks of each kind of approach. We make comparisons both theoretically and through numerical experiments.

We outline a procedure for breaking up the data and formulating the RF optimization problem for polytopal, ellipsoidal, and other types of fragmentations. We prove convergence results to demonstrate the consistency of our approach. RF can handle overlapping regions in the fragmentation by adapting the way statistics are computed. Our algorithm for fragmenting the data relies heavily on strategic clustering of the data sample to dissect the modality. We discuss methods and heuristics for implementing our approach. We extend and compare different clustering methods in terms of how well they serve as a preliminary step to our procedure. We consider applications in management science and financial mathematics and perform numerical experiments to demonstrate how the approach fits into existing frameworks and to evaluate its performance. It turns out that the new approach has a clear advantage when the data is multimodal and the training sample is small, but not too small relative to the number of modes. Additionally, we illustrate some extensions to the general model. The black swan approach involves adding an extra layer of robustness to account for rare and unexpected events of large magnitude and consequence. Two-stage and multistage stochastic programs extend the framework to decisions that must be made in stages, where after each stage random occurences influence the parameters of the next stage. We derive formulations for these extended models and investigate the production transportation-problem as an application.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Consider a decision maker who wants to optimize some outcome. She knows how her decision will affect the outcome, but she also knows that other factors will affect the outcome as well. These factors are outside of her control and unknown at the time of the decision. Such factors can be encoded in the form of parameters which are random variables. This problem structure covers many applications. For example, managers must choose inventory levels to maximize profit under uncertain demand, investors choose a portfolio to maximize gain when returns are random, and service providers must pick facility locations to minimize distances traveled where demand locations are unknown.

This type of decision making under uncertainty is a well studied problem in operations research, where one must optimize a known objective function over decision variables and unknown parameters. Many approaches can be taken to account for the unknown parameters, depending on the degree of uncertainty.

Stochastic programming is a classical framework for modeling optimization problems with uncertain parameters. Suppose one is interested in solving a problem such as we have described:

$$\min_{\mathbf{x}\in\mathscr{X}} \quad f(\mathbf{x}, \boldsymbol{\xi}),$$

where $\mathscr{X}$ is some convex feasible set and f is some convex cost function. However, the value of $\boldsymbol{\xi}$ is unknown at the time of the decision. For example, in inventory problems, customer demand is unknown when stock quantities are chosen. We represent the

uncertainty in $\boldsymbol{\xi}$ as a random variable, Ξ, where $\boldsymbol{\xi}$ denotes the realization of Ξ. The decision and realization together result in some outcome $f(\mathbf{x}, \boldsymbol{\xi}) \in \mathbb{R}$. The cost function f is known. If one chooses a value that will optimize the outcome in terms of expectation, then the problem becomes a stochastic program:

$$\min_{\mathbf{x} \in \mathscr{X}} \quad \mathbf{E}_{\pi}[f(\mathbf{x}, \Xi)],$$

where Ξ is a random vector $\Xi : D \to \Omega$ with realization $\boldsymbol{\xi}$, equipped with probability space $(D, \mathscr{D}, P)$ and measurable space $(\Omega, \mathscr{F})$. We consider $\Omega = \mathbb{R}^p$ and $\mathscr{F}$ as the Borel σ-algebra, so that π lies in the space of probability measures Θ over the measurable space $(\mathbb{R}^p, \mathscr{F})$. We refer the reader to Shapiro and Dentcheva (2014) for a thorough review of stochastic programming.

In theory, solving this stochastic program will lead to a solution that will perform best in terms of expectation. However, in practice, there are challenges to this approach beyond the computational difficulties of Monte Carlo approximations to multidimensional integrals as in Shapiro and Homem-de-Mello (2000). A bigger challenge is that knowledge about exact distributions may be unavailable or too costly. In fact, in practice it is rare that the distribution is known; imagine cases such as customer demand or investment returns. True distributions often do not exist or are subject to changes over time or contaminations through noise. Having exact knowledge of the distribution is quite unrealistic. Therefore, some kind of approximation will have to be made. We assume that distributions are unknown and only a set of historical data is available. Given some past realizations of Ξ, a decision maker must choose $\mathbf{x}$ in an attempt to minimize f without knowing its true value or expectation.

Given such a decision problem with unknown distribution and a set of historical data, many approaches can be taken to approximate the true problem. Parametric models assume a family of distributions for the data, requiring strong probabilistic assumptions. We aim to reduce model assumptions and so consider nonparametric approaches. For parametric models, the choice of parameters is determined by the model and only the values are fit to the data. For nonparametric, the choice of parameters is determined by the training data.

A natural approach would be to approximate the unknown distribution with the

empirical distribution and solve the resulting stochastic program. Using the empirical distribution as an approximation to the true distribution is equivalent to replacing the expectation in the objective function by a sample average, as follows:

$$\min_{\mathbf{x}\in\mathscr{X}} \quad \frac{1}{N}\sum_{j=1}^{N}[f(\mathbf{x},\boldsymbol{\xi}_j)], \tag{1.1}$$

which is referred to as a sample average approximation (SAA). The solution optimizes the objective function over the training set and thus may overfit to the training data. The propensity to fit to individual fluctuations rather than larger patterns can cause the method to perform poorly under moderate size samples, as illustrated in Bertsimas and Thiele (2006), or under noise or contamination, as in Hanasusanto et al. (2014), both for the newsvendor problem. The method tends not to be as robust and has high variance, as documented for portfolio optimization in Lim et al. (2011) and DeMiguel et al. (2009).

Another way of approximating the true stochastic program would be to replace the unknown distribution with a worst case distribution. This reflects an aversion to distributional ambiguity in the decision maker, which has been well defined and justified through decision theory and empirical observations, as in Gilboa and Schmeidler (1989) and Epstein (1999). This concept led Scarf to develop a robust version for the stochastic program which considers the worst-case expectation over a distribution set (1958). This technique is known as distributionally robust optimization (DRO), where the unknown true distribution is replaced with the worst-case distribution for each set of decision values. At this point we would like to mention that although the approach we will present later on also relies on distributional ambiguity, we do not assume that the decision maker is averse to distributional ambiguity; the use of our method is justified even if the decision maker is ambiguity neutral. In DRO, the solution may be biased but is less vulnerable to distributional ambiguity and often has lower variance (as we demonstrate for some examples). The distribution set is defined by constraints derived from prior information. The most commonly used set of constraints, and the ones we will consider, are the first and second moments. In the case where only historical data is available, the moments must be estimated. The formulation of the standard two moments constrained distributionally robust optimization problem (MDRO) is as

follows:

$$\begin{aligned} \min_{\mathbf{x}\in\mathscr{X}} \quad & \sup_{\pi\in\Theta} \ \mathbb{E}_\pi[f(\mathbf{x},\Xi)], \\ \text{subject to} \quad & \mathbb{E}_\pi[\Xi] = \boldsymbol{\mu}, \\ & \mathbb{E}_\pi[\Xi\Xi^T] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T, \\ & \mathbb{E}_\pi[I_{\mathbb{R}^p}] = 1, \end{aligned} \tag{1.2}$$

where π lies in the space of probability measures Θ over the measurable space $(\mathbb{R}^p, \mathscr{F})$. Although MDRO is robust to any distribution satisfying the moments constraints, it is often too conservative, as many qualifying distributions do not realistically reflect the data. For example, Scarf showed that in the case of the one-dimensional newsvendor problem, the extremal distribution is always a two-point distribution (1958). Other authors have shown that this result holds for more complicated objective functions: Zhu et al. (2006) and Yue et al. (2006) for the case of a newsvendor problem optimizing various types of relative regret and Popescu (2007) for so-called one or two-point support functions. Corresponding decisions, based on such unrealistic worst-case distributions, may not perform well in practice if the goal is to minimize expected loss. MDRO does not make very effective use of the data either by condensing it to just mean and covariance and discarding all other information. The method cannot distinguish between vastly different distributions, such as normal and exponential, with the same moments.

Since its genesis in the 1950s, the MDRO model has expanded to multiple dimensions and relies on duality theory for moments problems due to Isii (1960). In the years since, MDRO has become a widely accepted method for modeling uncertainty due to advances in conic optimization, such as in the newsvendor problem in Gallego and Moon (1993), and in the general case in Bertsimas et al. (2010), Dupacova (1966), Dupacova (1987), and Shapiro and Ahmed (2004), amongst many others. Weisemann et al. (2014) provide a generalized framework for DRO where the distributional robustness is modeled in the constraints. Other descriptive statistics beyond the first two moments, such as measures of symmetry, independence, and unimodality have been considered as well (Hu and Hong (2013), Popescu (2005)). Other authors consider general classes of mixed integer linear programs with uncertain objective coefficients (Li et al. (2014), Natarajan et al. (2011)) or linear optimization problems with uncertainties in both the

objective and constraints (Goh and Sim (2010)). Many other papers apply DRO to objective functions measuring various types of regret (Natarajan et al. (2013), Zhu et al. (2006), Perakis and Roels (2008), Yue et al. (2006)).

Delage and Ye (2010) consider a data-driven extension of MDRO. Their method allows for further ambiguity through moments uncertainty, enlarging the ambiguity set from classical MDRO.Knowledge of the support and a confidence region for the moments are the only requirements, which can be constructed from historical data. They achieve probability guarantees that the true distribution lies in the ambiguity set. As such, with more data they are able to reduce the size of confidence sets for the parameters. However, the ambiguity set is enlarged under their method, whereas we are interested in restricting it to reduce conservativeness.

We wish to introduce a data-driven approach that performs well in terms of minimizing expected loss and yet maintains robustness. We attempt to use the data to approximate the true stochastic program model in the most effective way possible. Similar to MDRO, our approach falls under the category of those that consider moments constraints. However, utilizing historical data, we divide the support into subregions and use conditional moments to constrain our ambiguity set in order to capture the modality of the data. We consider first and second order conditional moments on each region. We then consider the worst case expectation over the ambiguity set, as in classical distributionally robust optimization. We want to break down the data enough to make sure that the ambiguity set reflects the structure of the data. However, we want to avoid fragmenting it into too small of pieces, lest we leave ourselves vulnerable to overfitting. In effect, we are condensing the data by region and then accounting for any distribution that would result in the same reduction. We would like to condense the data such that we introduce some ambiguity without allowing for too many possibilities.

As we fragment the support of the training data to construct our ambiguity set, we term our method robust fragmentation (RF). We develop RF in an attempt to construct a robust, scalable, and effective framework for decision making. We combine modeling and fitting into one step. The parameters for our optimization model come directly from the data and inform the structure of our assumptions. The values, number, and type of parameters will depend on the shape and size of the data. The use of our method is justified by performance in terms of the expected loss even when there is no

noise, contamination, or changing distributions. We allow for two types of distributional ambiguity. The first type is what we call support ambiguity. This is when we allow for distributions that have different support, such as in MRDO. The second type is ambiguity in the probability measure. We allow for ambiguity in the probability vector whose components represent the probability of each subregion.

There have been a number of works that consider robust optimization in which a probability vector is treated as an uncertain parameter. In Wang et al. (2013), a likelihood robust optimization model is developed, where the ambiguity set consists of all distributions for which the data sample achieves a certain level of likelihood. This includes all probability vectors, over N distinct sample points with some sample frequencies, that have likelihood above some threshold, where the sample proportions maximizes the likelihood. They develop connections between their approach and Bayesian statistics and empirical likelihood theory. Their general setup assumes a discrete number of scenarios, but they additionally explore the possibility of continuous random variables in the one dimensional case by constructing a band around the cumulative distribution function. Ben-Tal et al. (2013) generalize this approach by using ϕ-divergences as a way to construct an ambiguity set of distributions around a nominal distribution (constructed from data). They show that uncertainty sets based on ϕ-divergences naturally arise when the parameter is a probability vector. In this case, expectations take the form of robust linear constraints in $\mathbf{p}$, where $\mathbf{p}$ is constrained by ϕ-divergence with the nominal distribution. They derive the dual of these constraints in terms of the conjugate function $\phi^\star$ of ϕ. They show that the dual of such a robust problem is tractable for many forms of ϕ in that it is a convex problem where the constraints admit a self-concordant barrier. We adopt their approach in terms of constructing an ambiguity set for a probability vector based on ϕ-divergences. However, in our case the vector represents the probability of each subregion, rather than of sample points. Additionally, we also combine this approach with ambiguity in the support over each subregion through moments constraints. We show that the problem is still tractable for many ϕ with the two layers of ambiguity.

Love and Bayraksan (2013) extend the concept of uncertainty regions defined by ϕ-divergences to two-stage stochastic programs. Both Ben-Tal et al. (2013) and Love and Bayraksan (2013) allow for ambiguity in the probability measure for each point,

but do not allow for support ambiguity. In Birbil et al. (2009), they consider a robust version of single-leg airline revenue management, where the distribution for the demand is assumed to be discrete, with the probabilities ambiguous over an ellipsoidal region. Robust optimization with ambiguity in parameters over ellipsoidal uncertainty sets was first introduced in Ben-Tal and Nemirovski (1998).

A few authors take other approaches to building a distributional ambiguity set that is constrained by a data sample. In Bertsimas et al. (2013), they develop robust sample average approximation, where the ambiguity set is defined by a goodness-of-fit hypothesis test. Thus, distributions that consist of small fluctuations from the empirical distribution are considered. They focus on asymptotic convergence and finite sample guarantees for their method. Mevissen et al. (2013) consider a data-driven version of distributionally robust polynomial optimization where the ambiguity set is defined as a norm ϵ ball around the empirical distribution, which must be solved using iterative SDP relaxations. Neither method directly addresses multimodality of distributions nor the link between sample data and support ambiguity.

There are many other recent data-driven methods that fall under the class of robust optimization (RO) (e.g. in Bertsimas et al. (2011)). Instead of considering probability distributions for the data, RO accounts for deterministic variability in the model parameters. Worst-case scenarios are considered over the set of all possible values for each parameter contained in some uncertainty set. See Ben-Tal et al. (2002) for a thorough review of robust optimization. RO can be viewed as a special case of DRO, when each distribution has all of its weight on a single point. Xu et al. (2012) show that every solution to a robust optimization problem is also a solution to a DRO problem. This equivalence allows them to construct RO formulations for sample-based problems that are statistically consistent. Bertsimas et al. (2013) construct uncertainty sets for linear robust optimization by using historical data and statistical tests to get finite sample probabilistic guarantees. They aim to leverage data to reduce the size of uncertainty sets while retaining robustness. While using data to constrain the problem, as we do, none of these methods consider distributional ambiguity or the probability of events occurring within uncertainty sets.

There are a host of other types of data-driven approaches to this type of problem, such as in Godfrey and Powell (2001) and Levi et al. (2007), which deal specifically

with the newsvendor model. Here we seek only to mention those that incorporate some measure of distributional ambiguity or other robustness.

In our proposed approach, we break up the support region into subregions and use conditional moments to define our ambiguity set. Similar partitioning approaches have been used before in specific contexts. Hanasusanto et al. (2014) try to adapt to multimodal demand distributions in the newsvendor problem by considering ellipsoidal regions. They develop a quadratic decision rule approximation and show that it achieves high accuracy in numerical tests. However, they consider only the newsvendor problem and assume prior knowledge of the support and conditional probabilities and moments, whereas we use the data to construct our fragmentation. Additionally, we consider other types of fragmentations besides ellipsoidal and frame our approach in terms of its relation to existing approaches. We introduce an additional layer of ambiguity in probability measure. We consider asymptotic convergence and implementation procedures as a data-driven method.

Natarajan et al. (2014) also consider a similar approach using second-order partitioned statistics, specifically for the newsvendor problem. They derive closed form solutions for the one-dimensional case where mean, variance, and semi variance are known. We include a closed-form solution for the RF newsvendor problem in $\mathbb{R}$ with two intervals. In the multivariate case, they consider partitioning the random variable above and below the mean vector and derive a computationally tractable lower bound. In this way, they are able to capture some information about asymmetry of the distribution. We consider different types of fragmentations that capture more detail such as the modality. We also generalize beyond the newsvendor problem, incorporate fragmentation of the support using data, and introduce ambiguity in probability measure.

Our main contribution is to introduce robust fragmentation as a robust data-driven approximation scheme to a stochastic program. The goal is to approximate the true, unknown stochastic program in an effective yet robust way. We seek to balance ambiguity in the support and in the probability measure while also tailoring to the structure, amount, and reliability of the data. We seek to position our approach as a generalization and unification of distributionally robust optimization and stochastic programming using the empirical distribution. Fragmentation allows us to dissect the data structure to better account for modality and reduce the problem size.

The method is especially applicable to distributions that are multimodal. If we partition intelligently, we have a greater level of detail than with using moments on the full support. There are many applications in which this is relevant. For example, for a given ordering cycle in an inventory problem, consider a seller with umbrellas, rain jackets, and sunglasses. The decision variable $\mathbf{x}$ is the inventories for the week and the demand is the random variable $\boldsymbol{\xi}$, where the objective is to maximize worst case expected profit. If it rains that week, demand for umbrellas and rain jackets will be high. However, if it's sunny, demand will be high for sunglasses. It may be both sunny and rainy during the week or neither. Also, more demand for umbrellas also lowers demand for rain jackets and vice versa as they are competing products. So there is a structure for the demand distribution that involves several modes that occur whether it is rainy, sunny, or both and may be dependent on the degree of the weather and other factors as well. In our application section, we will demonstrate that RF performs very well when the true generating distribution is multimodal.

The remainder of the paper is as follows. In Chapter 2, we formally introduce our approach. We show how to formulate the problem in a tractable way for a range of objective functions and for the univariate case. The multivariate case is included for many different fragmentation types, including ellipsoids, differences of ellipsoids, and polytopes. We also demonstrate asymptotic convergence based on results from Sun and Xu (2015). In Chapter 3, we discuss a special case where the fragmentation is not disjoint. By changing the procedure by which statistics are computed, we deal with this scenario. A procedure for general RF implementation is discussed in Chapter 4. An important preliminary step is to capture the modality of the data through clustering methods. In Chapter 5, we develop new and extend existing clustering techniques for this purpose and compare their benefits and drawbacks. In Chapter 6, we explore applications in the newsvendor model, portfolio selection, and facility location and present numerical results to demonstrate the concepts of our approach and exhibit performance. In Chapter 7, black swan events are considered. These are events with a monumental impact that have not been seen in the past. The goal is to try to provide a shield for their consequences. In Chapter 8, two-stage and multistage stochastic programs are used to extend the framework to decisions that must be made in stages. Random events or variables will be realized after each stage. We briefly conclude in Chapter 9. The

appendices include some preparatory material (A.1), details on an example (A.2), and a list of acronyms (B.1).

Chapter 2

Robust Fragmentation

We are concerned with decision making problems in which parameter distributions are unknown but some historical data is available. The decision maker must choose how to efficiently and robustly use such data to approximate the true stochastic program in an attempt to optimize the expected outcome. We develop robust fragmentation (RF) in order to construct ambiguity sets that are meaningful, statistically consistent with historical data, and result in tractable formulations. In RF, we fragment the data containing past realizations of Ξ into pieces to dissect its structure. We use regional probabilities and first and second order conditional moments on each subregion containing a subsample to construct an ambiguity set of distributions. Thus, the parameters that govern our ambiguity set come directly from the data and inform the structure of our assumptions. We consider two layers of distributional ambiguity: support ambiguity and probability ambiguity.

In order to construct uncertainty sets describing ambiguity in probability, we will use ϕ-divergences. The ϕ-divergence between two M-dimensional vectors $\mathbf{p}, \mathbf{q} \geq 0$ is given by

$$I_\phi(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^{K} q_k \phi\left(\frac{p_k}{q_k}\right),$$

where $\phi(t)$ is convex for $t \geq 0$, $\phi(1) = 0$, $0\phi(\frac{a}{0}) = a \lim_{t\to\infty} \frac{\phi(t)}{t}$ for $a > 0$, and $0\phi(\frac{0}{0}) = 0$. For probability vectors, we have the additional constraints that $\mathbf{e}^T\mathbf{p} = 1$ and $\mathbf{e}^T\mathbf{q} = 1$.

Some commonly used ϕ-divergence functions include $\phi(t) = t\log(t) - t + 1$ (Kullback-Leibler) and $\phi(t) = \frac{1}{t}(t-1)^2$ (χ^2 distance). For more on ϕ-divergences, see Pardo (2006). The conjugate of ϕ is a function $\phi^\star : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ defined as $\phi^\star(s) = \sup_{t \geq 0}\{st - \phi(t)\}$. We will see that robust constraints in terms of probability vectors can be reformulated in terms of conjugate functions. As examples, the conjugate functions of the Kullback-Leibler divergence and χ^2 distance are $\phi^\star(s) = e^s - 1$ and $\phi^\star(s) = 2 - 2\sqrt{1-s}$ for $s < 1$, respectively. It is not always the case that $\phi^\star$ exists in closed form. In some of these cases it may still be possible to determine tractable reformulations of the robust problem, but we restrict our consideration to cases where the closed form exists.

We follow Ben-Tal et al. (2013) in terms of constructing an uncertainty region $\mathscr{P}$ for probability vector p based on ϕ-divergence. We consider the case where there are K support regions for random variable Ξ. The vector $\mathbf{p}$ represents the probability that any of K different scenarios will occur. Scenario k is that $\xi \in \Omega_k$, so that p_k represents the probability that a realization of Ξ will lie in Ω_k. We have a set of historical data, that is, realizations of Ξ. Each ξ_i, $i = 1, ..., N$ is i.i.d. and thus we can compute estimated probabilities $\hat{p}_k = \frac{1}{N}\sum_{i=1}^{N} I(\xi_i \in \Omega_k)$. We consider a discrete random variable Z with a fixed sample space $\{1, ..., K\}$, where $Z = k$ if $\xi \in \Omega_k$ and thus $\mathbb{P}(Z = k) = p_k$. We denote the density function of Z as $f_{\mathbf{p}}$, which depends on the parameter $\mathbf{p}$. Under the assumption that a random sample of size N was drawn from Z, the ϕ-divergence between the true density $f_{\mathbf{p}}$ and the estimated density $f_{\hat{\mathbf{p}}}$ is given by

$$I_\phi(f_{\mathbf{p}}, f_{\hat{\mathbf{p}}}) = \int f_{\hat{\mathbf{p}}}\phi\left(\frac{f_{\mathbf{p}}}{f_{\hat{\mathbf{p}}}}\right) d\mu = \sum_{k=1}^{K} \hat{p}_k \phi\left(\frac{p_k}{\hat{p}_k}\right), \tag{2.1}$$

where μ is the counting measure. We follow Pardo (2006) in order to construct our uncertainty set for $\mathbf{p}$. We assume that ϕ is twice continuously differentiable in a neighborhood of 1 with $\phi''(1) > 0$. Under some regularity conditions, Pardo (2006) shows that $\frac{2N}{\phi''(1)} I_\phi(f_{\mathbf{p}}, f_{\hat{\mathbf{p}}})$ converges in distribution to a χ^2 distribution with $K - 1$ degrees of freedom. Thus, if we set $\rho = \frac{\phi''(1)}{2N}\chi^2_{K-1,1-\alpha}$, then the uncertainty set $\mathscr{P} = \{\mathbf{p} \in \mathbb{R}^K | \mathbf{p} \geq 0, \mathbf{e}^T\mathbf{p} = 1, I_\phi(\mathbf{p}, \hat{\mathbf{p}}) \leq \rho\}$ is an approximate confidence set of confidence level $1 - \alpha$. Since the confidence set is based on asymptotic results, we may add correction terms to ensure better coverage for small sample sizes, as described in

Pardo (2006). If we set $\rho = 0$, then we must have that $p_k = \hat{p}_k \ \forall k$, since $\phi(1) = 0$ and $\phi(x) \neq 0$ for all other x. In this case $\mathscr{P} = \{\mathbf{p}\}$ and there is no probability ambiguity.

In addition to constructing an ambiguity set for the probability p_k of each subregion through the estimated probabilities $\hat{p}_k$, we allow for ambiguity in the support inside each region. Over each subregion, we compute the conditional mean $\hat{\boldsymbol{\mu}}_k$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_k$ of the corresponding subsample, and use these to constrain an ambiguity set Θ of distributions. We then consider the worst case expectation as in distributionally robust optimization:

$$\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X}} \quad & \sup_{\pi\in\Theta} \ \mathbb{E}_\pi[f(\mathbf{x},\Xi)], \\
\text{subject to} \quad & \mathbb{P}_\pi[\Xi \in \Omega_k] = p_k, \\
& \mathbb{E}_\pi[\Xi|\Xi \in \Omega_k] = \hat{\boldsymbol{\mu}}_k, \\
& \mathbb{E}_\pi[\Xi\Xi^T|\Xi \in \Omega_k] = \hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T, \quad \text{for } k = 1, ..., K, \\
& \mathbf{e}^T\mathbf{p} = 1, \\
& \mathbf{p} \geq 0, \\
& \sum_{k=1}^{K} \hat{p}_k \phi(\frac{p_k}{\hat{p}_k}) \leq \rho,
\end{aligned} \tag{2.2}$$

where each region $\Omega_k \subseteq \Omega$, and π_k lies in the space of probability measures Θ_k over the measurable space $(\Omega_k, \mathscr{F})$. For now, we assume that $\Omega_k \cap \Omega_{k'} = \emptyset$ for any k, k'. We will deal with the overlapping case in Chapter 3. In general, it need not be the case that $\cup\Omega_k = \mathbb{R}^p$. The set of moments constraints must be feasible. In Section 2.1, we will discuss how to reformulate the problem. We assume that the constraints $\mathbf{x} \in \mathscr{X}$ are linear constraints.

Assuming the regions Ω_k are disjoint, we can rewrite the objective function as

$$\min_{\mathbf{x}\in\mathscr{X}} \quad \sup_{\mathbf{p}\in\mathscr{P}} \quad \sum_{k=1}^{K} p_k \sup_{\pi_k\in\Theta_k} \mathbb{E}_{\pi_k}[f(\mathbf{x},\Xi)],$$

where π_k represents the conditional distribution for Ξ over Ω_k. In this representation it is easy to see the two layers of ambiguity. We aim to achieve a balance between these two types of ambiguity, in support and probability. If we allow for maximal support

ambiguity, as in MDRO, then there is no probability ambiguity. As we break up the support into subregions, we restrict the support, but we cannot be certain about how probable each subregion is. On the extreme end, if we assume exact support points, then we cannot be confident about their probabilities of occurring with a limited amount of data.

In effect, we are constructing an ambiguity set of mixture distributions with probabilities of selection $\mathbf{p}$ and components π_k, which have restricted support Ω_k. For each component in the mixture, we are considering an ambiguity set constructed from the mean and covariance of the sample in that region of the support. Additionally, we are considering an ambiguity set for the mixture weights $\mathbf{p}$.

The process of obtaining conditional moments from historical data can be thought of as condensing the data regionally and then accounting for any distribution that would result in the same reduction. We want to condense the data enough to avoid overfitting, without condensing too much and causing the ambiguity set to not reflect the structure of the data. Conversely, our approach can be understood as a fragmentation of the data into subsamples in a way that best captures the modal structure of the data. We then use this modal structure to construct our ambiguity set.

Our framework includes classic distributionally robust optimization with moments constraints (MDRO) as a special case. MDRO, as in (1.2), clearly falls under (2.2) with $K = 1$, $p = 1$ (no probability ambiguity), and $\Omega_1 = \mathbb{R}^p$. No fragmentation occurs and the data is fully condensed into overall moments and a conservative decision is made based on the worst-case expectation. All of the ambiguity is in the support of the distribution, which is unrestricted as long as the moments constraints are satisfied. As discussed in the introduction, although MDRO accounts well for distributional ambiguity, the solution may be biased, conservative, and not perform well in practice if the objective is to minimize expected loss (not considering ambiguity aversion). The ambiguity set does not capture the structure of the data, because no fragmentation has occurred. The ambiguity set contains no information about the probability of Ξ occuring in different parts of the support.

Stochastic programming using the empirical distribution (SAA), as in (1.1), also falls under (2.2) as well, with $K = N$, $\hat{\boldsymbol{\mu}}_k = \boldsymbol{\xi}_k$, $\hat{\boldsymbol{\Sigma}}_k = \mathbf{0}$, $\rho = 0$ and $\hat{p}_k = \frac{1}{N}$ for $k = 1, ..., N$. It lies on the other extreme, where the data is separated completely into individual

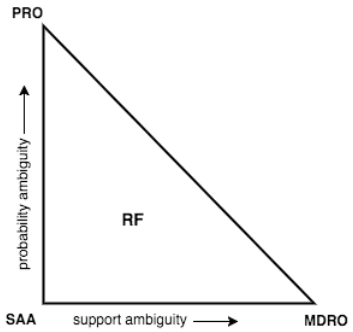

Figure 2.1: RF as a generalization

records, restricting to a single distribution (empirical). The data is not condensed at all and there is no support ambiguity. As described in the introduction, SAA may overfit to the training data and performs poorly with small samples or under changing or noisy conditions. The data can be fragmented too much, fitting to individual fluctuations rather than larger patterns. SAA can be made robust by adding probability ambiguity, as we discussed and as done in Ben-Tal et al. (2013). The robust version falls under (2.2) with the same structure as standard SAA but with $\rho > 0$ defining an uncertainty set for $\mathbf{p}$. We will refer to this robust problem, with an uncertainty set for $\mathbf{p}$ but no support ambiguity (each region is a singleton), as probability robust optimization (PRO). Such a model only accounts for ambiguity in probability and not in support. This does not eliminate the problem of overfitting, especially for small N, since the location of the sample points completely determines the support for the set of feasible distributions.

The goal of RF is to allow for different balances of ambiguity in probability and support in order to adapt to the amount and structure of historical data available. RF is a generalization of MDRO, PRO, and SAA as depicted in Figure 2.1. The extent of the fragmentation acts as a tuning parameter that exchanges ambiguity in support for ambiguity in probability. We argue that in most cases, a middle ground will be the most generalizable.

We explore the following simple example in order to help illustrate some of the fundamental concepts and uses of RF. Suppose we consider a one-dimensional problem, with small sample size N. In Figure 2.2, we show a data sample generated from a bimodal distribution and a selected distribution from each ambiguity set. The SAA

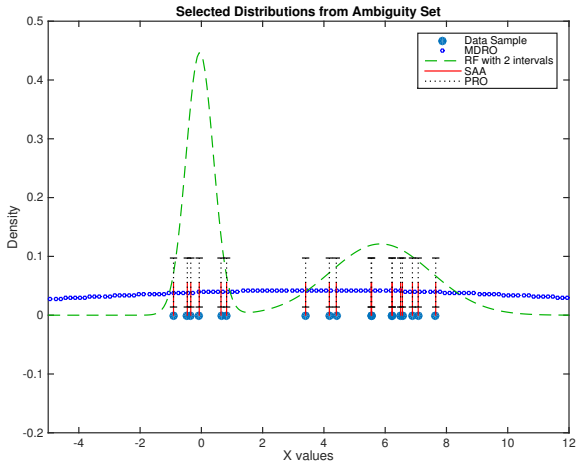

Figure 2.2: Distributions from Ambiguity Sets

considers only the empirical distribution in red; PRO considers different weights for the sample points that satisfy the ϕ-divergence constraint as depicted in purple; MDRO considers any distribution with the same moments such as the normal distribution in blue; RF considers any distribution with the proper modal characteristics such as the one in green. From this simple example, we can see how MDRO may be too ambiguous, as the blue density does not fit the data well and doesn't account for the modality. We can also see that SAA and PRO may rely too much on the specific sample for small N. Of course, the MDRO ambiguity set also includes the green density as well as the red, but the important point is that it also accounts for many distributions that don't follow the structure of the data and that are removed with RF.

Let's consider an even simpler example. Suppose we have a discrete distribution for the one-dimensional newsvendor problem, where a manager must determine an inventory level x to minimize a cost $|x - \Xi|$ for the resulting overstock or understock, where Ξ is the unknown demand. Suppose the distribution is unknown but that we have a historical sample of size N. For a given sample, we can compute the SAA and MDRO solutions, and specific cases of PRO and RF solutions (see Appendix A.2 and Theorem (7)). For RF, we consider a fragmentation into two unbounded intervals, which we will refer to as RF-2. For each of the four solutions $\hat{x}_N$, we can calculate the expected cost over a sample of size N, $\mathbb{E}[|\hat{x}_N - \Xi|]$, where the expectation is over both the random

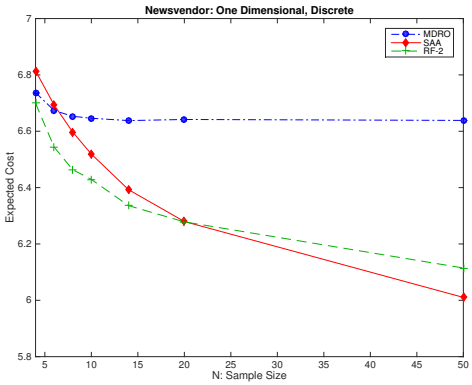

Figure 2.3: Support Ambiguity

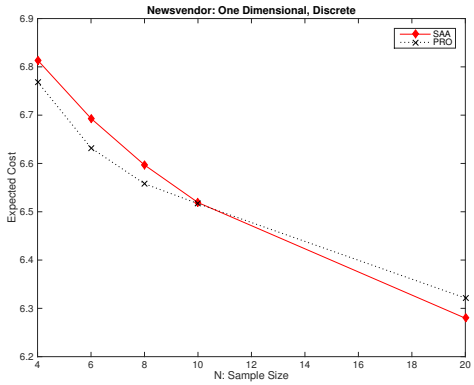

Figure 2.4: Probability Ambiguity

estimator $\hat{x}_N$ and the random demand Ξ. As the objective is to minimize the expected cost $\mathbb{E}[|x - \Xi|]$ we can directly compare the quality of the solutions. We compute these values for selected N and plot them in Figures 2.3 and 2.4. Details are given in *Example 1* in Appendix A.2.

On the left (Figure 2.3), we plot the MDRO, SAA, and RF-2 solutions to illustrate the effects of changing the support ambiguity. The results show that in this case, MDRO is best for very small N, with SAA most effective for large N. RF-2 is the best method for intermediate values of N. Keep in mind that this is only one choice of fragmentation for RF, a very simple one with no probability ambiguity. Other fragmentations would likely perform even better. The optimal RF would be at least as good as RF-2, which we chose for the simplicity of a closed form solution. The trend of RF performing best for intermediate N is one that we will explore in more depth later on, but is very intuitive. As N increases, we can afford to have more complex models (more parameters). For very small sample size, all parameter estimates will have high variance and so additional regional estimates will only add more estimation error without gaining much information about the data modality. For very large sample size, the empirical distribution will be close to the true distribution and so SAA will be difficult to outperform (although it may be more computationally expensive). In fact, the SAA solution will always be a consistent estimator of the true solution. For this example, the sample median $\hat{m}$ is a consistent estimator for the median. We can design a fragmentation with more regions to outperform SAA as N increases; we will discuss how to choose the number

and location of regions in Chapters 4 and 5. The MDRO solution, the sample mean, is not consistent for the median as the distribution is asymmetric. However, it can be shown to have smaller variance, which is why it performs better for very small N. The performance of RF for intermediate N relies on this consistency-variance tradeoff. RF may dissect modal information that improves it over MDRO and condense the data to improve upon SAA. The plot on the right (Figure 2.4) illustrates the effect of changing the ambiguity in probability. We can see that for small N, PRO outperforms SAA. The ambiguity helps to avoid overfitting. In general, we may combine both types of ambiguity and tailor them according to the size and structure of the data sample.

Besides a generalization of existing nonparametric approaches such as MDRO and SAA, we describe some other angles from which to view RF. RF can be employed as an adaptive procedure that adjusts as data comes in. As the size and structure of the available data changes, it will be advantageous to update the strategy accordingly. As N increases, we can afford to build more complex models.

RF can also be framed in terms of a comparison to SAA. RF, instead of using the empirical distribution as an approximation to the truth, reduces the problem size by summarizing the data regionally. Noise is averaged out and the training sample is replaced with a more manageable set of parameters. This increases computational efficiency and reduces variability in the solution. RF can operate effectively on any size sample; we simply adjust the fragmentation accordingly. SAA requires a large training sample with little noise to be effective. RF may be more robust to noisy, changing, and limited data. In some cases, the data to be applied on may not even come from the same distribution as the training data or there may not be a true generating distribution. We want to prevent overfitting to past and possibly less relevant data. Under any of these cases, it is to our benefit to condense the data and allow for some ambiguity.

RF can be also thought of as an extension of MDRO, where the support is divided and conditional moments are used rather than full moments. RF reduces conservativeness by restricting the ambiguity set to more closely resemble the true distribution, given sufficient data. RF will capture much greater detail from the data. For example, if the data is multimodal, breaking it up into pieces will capture the modality. For example, in retail, fashion trends may dictate a popularity state for each product, with different states having vastly different distributions that are more accurately described

by regional moments, as in Hanasusanto et al. (2014).

2.1 RF Dual Formulation

Let's return to the RF formulation as in (2.2). We will assume that $\cap int(\Omega_k) = \emptyset$ for now and address the issue of overlapping regions in Chapter 3. Thus, $\mathbf{e}^T\mathbf{p} = 1$. Some of the regions Ω_k may be point masses at $\mathbf{m}_k$, with mass p_k and zero covariance. They can be pulled out of the expectation. Given that the regions are disjoint, the expectation may be broken into the sum of conditional expectations. Let S be the set of all indices k such that Ω_k is a point mass and C be the set of all other indices. In this case, an equivalent formulation to (2.2) is as follows:

$$\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X}} \quad & \sup_{\mathbf{p}\in\mathscr{P}} \sum_{k\in C} p_k \sup_{\pi_k\in\Theta_k} \int_{\Omega_k} f(\mathbf{x},\boldsymbol{\xi}) d\pi_k(\boldsymbol{\xi}) + \sum_{k\in S} p_k f(\mathbf{x},\mathbf{m}_k), \\
\text{subject to} \quad & \mathbf{e}^T\mathbf{p} = 1, \\
& \mathbf{p} \geq 0, \\
& \sum_{k=1}^{K} \hat{p}_k \phi(\frac{p_k}{\hat{p}_k}) \leq \rho, \\
& \int_{\Omega_k} d\pi_k(\boldsymbol{\xi}) = 1, \\
& \int_{\Omega_k} \boldsymbol{\xi} d\pi_k(\boldsymbol{\xi}) = \hat{\boldsymbol{\mu}}_k, \\
& \int_{\Omega_k} \boldsymbol{\xi}\boldsymbol{\xi}^T d\pi_k(\boldsymbol{\xi}) = \hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T, \ \forall k = 1,...,r,
\end{aligned} \tag{2.3}$$

where π_k is the conditional distribution for $\boldsymbol{\xi}$ given that it is in Ω_k. From this point forward, we will always assume the indices have been reordered such that $k \in C = \{1,...,r\}$ and $k \in S = \{k+1,...,K\}$.

In theory, this formulation could be considered with the assumption that conditional probabilities and moments were known exactly. However, we wish to develop a data-driven method that can be used in real applications and so assume parameters must be estimated from a data sample. Therefore, any region that contains only one point from the data sample will be equivalent to a point mass (estimated mean at the point,

estimated covariance of zero). Thus, as K increases and we shrink the size of our regions Ω_k, regions will transition to point masses. As K increases, regional integral terms in the objective function will move into the summation and the support ambiguity will shrink. If there is no probability ambiguity ($\rho = 0$), this illustrates the transformation from MDRO to SAA. If we introduce probability ambiguity ($\rho > 0$), we transition to PRO. As K increases, more probability ambiguity is introduced as the approximate $1-\alpha$ confidence set for $\mathbf{p}$ relies on the upper bound $\rho = \frac{\phi''(1)}{2N}\chi^2_{K-1,1-\alpha}$ on the ϕ-divergence, which is increasing in K (Pardo (2006)). We could, of course, adjust the value of ρ to change the level of probability ambiguity for any K.

In order to be able to reformulate (2.3) in a tractable way, we must restrict the class of cost functions f and ϕ-divergence functions ϕ that we consider. We must have that f is piecewise linear or quadratic and convex.

Assumption 1. *Suppose the cost function can be written as* $f(\mathbf{x}, \boldsymbol{\xi}) = \max_{l=1,...,L} \; f_l(\mathbf{x}, \boldsymbol{\xi})$*, where* $f_l(\mathbf{x}, \boldsymbol{\xi}) = \boldsymbol{\xi}^T \mathbf{A}_l \boldsymbol{\xi} + \mathbf{b}_l^T \boldsymbol{\xi} + c_l + \boldsymbol{\beta}_l^T \mathbf{x} + \mathbf{x}^T \mathbf{D}_l \boldsymbol{\xi}$ *and* $\mathbf{A}_l$, $\mathbf{b}_l$, c_l, $\boldsymbol{\beta}_l$, *and* $\mathbf{D}_l$ *are known parameters* $\forall l = 1, ..., L$.

Note that this structure includes many well known cost functions such as for the newsvendor problem, portfolio investment, conditional value-at-risk, L_1 distance and L_2 distance squared, amongst many others. We also must restrict the class of ϕ-divergences that we consider.

Assumption 2. *Suppose ϕ is a ϕ-divergence function s.t. the conjugate $\phi^\star$ exists in closed form and is monotone increasing.*

Note that Assumption 2 is met for many common ϕ-divergence functions such as Kullback-Leibler, χ^2-distance, and variation distance.

Theorem 1. *Under Assumptions* (1) *and* (2)*, we can formulate the dual of* (2.3) *as*

the following convex minimization problem:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{\gamma_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \eta+\rho\lambda+\lambda\sum_{k=1}^{r}\hat{p}_k\phi^\star\left(\frac{z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta}{\lambda}\right)\\
& +\lambda\sum_{k=r+1}^{K}\hat{p}_k\phi^\star\left(\frac{\gamma_k-\eta}{\lambda}\right),\\
\textit{subject to} \quad & \begin{pmatrix}1 & \boldsymbol{\xi}^T\end{pmatrix}\begin{pmatrix}z_k-c_l-\boldsymbol{\beta}_l^T\mathbf{x} & \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x})^T\\ \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x}) & \mathbf{Z}_k-\mathbf{A}_l\end{pmatrix}\begin{pmatrix}1\\ \boldsymbol{\xi}\end{pmatrix}\geq 0\\
& \forall\boldsymbol{\xi}\in\Omega_k,\forall l=1,...,L,\ \forall k=1,...,r,\\
& z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\\
& \forall k=1,...,r,\\
& \gamma_k\geq(\boldsymbol{\beta}_l+\mathbf{D}_l\mathbf{m}_k)^T\mathbf{x}+\mathbf{m}_k^T\mathbf{A}_l\mathbf{m}_k+\mathbf{b}_l^T\mathbf{m}_k+c_l,\ \forall\, l=1,...,L,\\
& \forall k=r+1,...,K,\\
& \gamma_k-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\ \forall k=r+1,...,K,\\
& \lambda\geq 0.
\end{aligned}
\tag{2.4}
$$

In the special case where there is no probability ambiguity ($\rho=0$), (2.4) simplifies to:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\{\gamma_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \sum_{k=1}^{r}\hat{p}_k\left(z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)\right)+\sum_{k=r+1}^{K}\hat{p}_k\gamma_k,\\
\textit{subject to} \quad & \begin{pmatrix}1 & \boldsymbol{\xi}^T\end{pmatrix}\begin{pmatrix}z_k-c_l-\boldsymbol{\beta}_l^T\mathbf{x} & \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x})^T\\ \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x}) & \mathbf{Z}_k-\mathbf{A}_l\end{pmatrix}\begin{pmatrix}1\\ \boldsymbol{\xi}\end{pmatrix}\geq 0,\\
& \forall\boldsymbol{\xi}\in\Omega_k,\forall l=1,...,L,\ \forall k=1,...,r,\\
& \gamma_k\geq(\boldsymbol{\beta}_l+\mathbf{D}_l\mathbf{m}_k)^T\mathbf{x}+\mathbf{m}_k^T\mathbf{A}_l\mathbf{m}_k+\mathbf{b}_l^T\mathbf{m}_k+c_l,\ \ \forall\, l=1,...,L,\\
& \forall k=r+1,...,K.
\end{aligned}
\tag{2.5}
$$

Proof. The proof is given in Appendix B.1. □

We should mention that in the remainder of this thesis we will present our theorems and propositions for the general case of $\rho > 0$, but that for the special case with no probability ambiguity, the formulations are identical, except with the simplification of the objective function and removal of some dual variables and constraints, as we see in (2.5). We will include these special cases as well, but they will not be the focus.

Starting with the simplest type of fragmentation with only one region (and thus no probability ambiguity), we would have the form in (2.5). For $\Omega_1 = \mathbb{R}_p$, this would be an SDP. An SDP is a linear optimization problem where all of the constraints are either linear or linear matrix inequalities (requiring positive definiteness of a matrix, all of whose entries are linear functions). This holds additionally for $\Omega = \mathbb{R}^+_p$ with $p \leq 3$ (due to Lemma 5 in Appendix A.1). For $\Omega_1 = \mathbb{R}^+_p$, it is in general an NP hard copositive program (see Dur (2010) for a survey on copositive programming). For RF, we are interested in cases where $r > 1$ and $\Omega_k \neq \mathbb{R}^p$, and so we expect the problem to be even harder in general. Our goal is to develop fragmentation structures that will lead to tractable versions of (2.4).

From this formulation it is clear that one major difficulty lies in the infinite dimensional constraints, and tractability will stem from an alternative representation of nonnegativity of a quadratic function over a subdomain Ω_k. We point out that if probability ambiguity is left out and only the conditional moments constraints are used, then the objective function reduces to linear and the constraints are convex. The other difficulty relates to the ϕ-divergence function. We need ϕ to be a function such that the objective is convex and such that if embedded in the constraints, the constraint set admits an explicit, self-concordant barrier function. Then the resulting convex problem can be solved by polynomial-time interior point algorithms.

Our approach will be to design fragmentations such that the quadratic inequalities become tractable in the form of LMIs, enabling us to reformulate (2.5) as an SDP, or (2.4) as an SCP (Self-Concordant Program, namely constrained convex minimization problem with a known self-concordant barrier function). We consider an SDP to be tractable due to its proven polynomial complexity and the existence of effective interior point method solvers such as SeDuMi and SDPT3 implemented in various software. We also consider an SCP to be tractable, to a lesser degree, due to the existence of polynomial-time interior point algorithms to solve it.

2.2 Univariate

We start by considering the one-dimensional case for (2.4). Here the constraints amount to requiring nonnegativity of a quadratic function over a region Ω_k. If we choose Ω_k such that it can be defined by a quadratic inequality, we can use the S-Lemma (see Lemma 3) to represent this as an LMI (for a survey of the S-Lemma, see Polik and Terlaky (2007)). Thus, we fragment $\mathbb{R}$ into a finite number of intervals $\Omega_k = [l_k, u_k]$ for $k \in C$ and point masses for $k \in S$. Let $a_k = l_k + u_k$ for any k. We may have intervals $(-\infty, u_1]$ and $[l_r, \infty)$ on the ends. A similar setup has been considered by Natarajan et al. (2014) in the specific context of the newsvendor problem.

Proposition 1. *Suppose $\xi \in \mathbb{R}$, the set of $\{\Omega_k\}$ consists of a finite number of interval regions r, of the form $[l_k, u_k]$, $(-\infty, u_1]$, or $[l_r, \infty)$, and point masses $m_1, ..., m_s$, and Assumptions 1 and 2 hold. Then the solution to* (2.2) *is given by the minimizer of the*

following convex problem:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{\gamma_k,z_k,\zeta_k,\upsilon_k,\tau_k\}} \quad & \eta+\rho\lambda+\lambda\sum_{k=1}^{r}\hat{p}_k\phi^{\star}\left(\frac{z_k+\zeta_k\hat{\mu}_k+\upsilon_k(\hat{\sigma}_k^2+\hat{\mu}_k^2)-\eta}{\lambda}\right)\\
& +\lambda\sum_{k=r+1}^{K}\hat{p}_k\phi^{\star}\left(\frac{\gamma_k-\eta}{\lambda}\right),\\
\textit{subject to } & \begin{pmatrix} z_k-c_l-\boldsymbol{\beta}_l^T\mathbf{x}-u_kl_k\tau_k & \frac{1}{2}(\zeta_k-b_l-\mathbf{D}_l^T\mathbf{x}+a_k\tau_k)\\ \frac{1}{2}(\zeta_k-b_l-\mathbf{D}_l^T\mathbf{x}+a_k\tau_k) & \upsilon_k-a_l-\tau_k\end{pmatrix}\\
& \succeq 0,\\
& \forall\, l=1,...,L,\ \forall\, k=2,...,r-1,\\
& \begin{pmatrix} z_r-c_l-\boldsymbol{\beta}_l^T\mathbf{x}+l_r\tau_r & \frac{1}{2}(\zeta_r-b_l-\mathbf{D}_l^T\mathbf{x}-\tau_r)\\ \frac{1}{2}(\zeta_r-b_l-\mathbf{D}_l^T\mathbf{x}-\tau_r) & \upsilon_r-a_l\end{pmatrix}\succeq 0,\\
& \forall\, l=1,...,L,\\
& \begin{pmatrix} z_1-c_l-\boldsymbol{\beta}_l^T\mathbf{x}-u_1\tau_1 & \frac{1}{2}(\zeta_1-b_l-\mathbf{D}_l^T\mathbf{x}+\tau_1)\\ \frac{1}{2}(\zeta_1-b_l-\mathbf{D}_l^T\mathbf{x}+\tau_1) & \upsilon_1-a_l\end{pmatrix}\succeq 0,\\
& \forall\, l=1,...,L,\\
& \tau_k\geq 0,\ \forall k=1,...,r,\\
& z_k+\zeta_k\hat{\mu}_k+\upsilon_k(\hat{\sigma}_k^2+\hat{\mu}_k^2)-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\ \forall k=1,...,r,\\
& \gamma_k\geq(\boldsymbol{\beta}_l+m_k\mathbf{D}_l)^T\mathbf{x}+a_lm_k^2+b_lm_k+c_l,\ \forall\, l=1,...,L,\\
& \forall\, k=r+1,...,K,\\
& \gamma_k-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\ \forall k=r+1,...,K,\\
& \lambda\geq 0.
\end{aligned}
\tag{2.6}
$$

Proof. From Proposition 1, we have that the general form of the dual is as in (2.4). The rL infinite dimensional constraints can be reformulated using S-Lemma (3). For an arbitrary internal interval $[l_k, u_k]$, for each l we have the constraint

$$\begin{pmatrix}1 & \xi\end{pmatrix}\begin{pmatrix} z_k - c_l - \boldsymbol{\beta}_l^T\mathbf{x} & \frac{1}{2}(\zeta_k - b_l - \mathbf{D}_l^T\mathbf{x}) \\ \frac{1}{2}(\zeta_k - b_l - \mathbf{D}_l^T\mathbf{x}) & \upsilon_k - a_l \end{pmatrix}\begin{pmatrix}1 \\ \xi\end{pmatrix} \geq 0 \text{ for all } l_k \leq \xi \leq u_k.$$

The interval condition may be restated as the quadratic inequality $(\xi - l_k)(\xi - u_k) \geq 0$. We can then employ the S-Lemma to reformulate the constraint as

$$\begin{pmatrix}1 & \xi\end{pmatrix}\left[\begin{pmatrix} z_k - c_l - \boldsymbol{\beta}_l^T\mathbf{x} & \frac{1}{2}(\zeta_k - b_l - \mathbf{D}_l^T\mathbf{x}) \\ \frac{1}{2}(\zeta_k - b_l - \mathbf{D}_l^T\mathbf{x}) & \upsilon_k - a_l \end{pmatrix} - \tau_k\begin{pmatrix} u_k l_k & -\frac{1}{2}(l_k + u_k) \\ -\frac{1}{2}(l_k + u_k) & 1 \end{pmatrix}\right]$$

$$\begin{pmatrix}1 \\ \xi\end{pmatrix} \geq 0 \ \forall \xi.$$ Similarly, for an unbounded interval such as (WLOG) $\Omega_r = [l_r, \infty)$, the constraint can be restated as $\xi - l_r \geq 0$, thus we can apply S-Lemma as well to get

$$\begin{pmatrix}1 & \xi\end{pmatrix}\left[\begin{pmatrix} z_k - c_l - \boldsymbol{\beta}_l^T\mathbf{x} & \frac{1}{2}(\zeta_k - b_l - \mathbf{D}_l^T\mathbf{x}) \\ \frac{1}{2}(\zeta_k - b_l - \mathbf{D}_l^T\mathbf{x}) & \upsilon_k - a_l \end{pmatrix} - \tau_r\begin{pmatrix} -l_r & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}\right]\begin{pmatrix}1 \\ \xi\end{pmatrix} \geq 0 \ \forall \xi.$$

If it is part of our fragmentation, we get a similar constraint for the unbounded interval on the other side. Thus, we get the dual formulation as in (2.6). □

In the special case of no probability ambiguity ($\rho = 0$), we have, by Proposition 1, that the general form of the dual is as in (2.5). Thus, we get the following formulation:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{\gamma_k,z_k,\zeta_k,\upsilon_k,\tau_k\}} \quad & \sum_{k=1}^{r}\hat{p}_k\left(z_k+\zeta_k\hat{\mu}_k+\upsilon_k(\hat{\sigma}_k^2+\hat{\mu}_k^2)\right)+\sum_{k=r+1}^{K}\hat{p}_k\gamma_k, \\
\text{subject to } & \begin{pmatrix} z_k-c_l-\boldsymbol{\beta}_l^T\mathbf{x}-u_kl_k\tau_k & \frac{1}{2}(\zeta_k-b_l-\mathbf{D}_l^T\mathbf{x}+a_k\tau_k) \\ \frac{1}{2}(\zeta_k-b_l-\mathbf{D}_l^T\mathbf{x}+a_k\tau_k) & \upsilon_k-a_l-\tau_k \end{pmatrix} \\
& \succeq 0, \\
& \forall\, l=1,...,L,\ \forall\, k=2,...,r-1, \\
& \begin{pmatrix} z_r-c_l-\boldsymbol{\beta}_l^T\mathbf{x}+l_r\tau_r & \frac{1}{2}(\zeta_r-b_l-\mathbf{D}_l^T\mathbf{x}-\tau_r) \\ \frac{1}{2}(\zeta_r-b_l-\mathbf{D}_l^T\mathbf{x}-\tau_r) & \upsilon_r-a_l \end{pmatrix} \succeq 0, \\
& \forall\, l=1,...,L, \\
& \begin{pmatrix} z_1-c_l-\boldsymbol{\beta}_l^T\mathbf{x}-u_1\tau_1 & \frac{1}{2}(\zeta_1-b_l-\mathbf{D}_l^T\mathbf{x}+\tau_1) \\ \frac{1}{2}(\zeta_1-b_l-\mathbf{D}_l^T\mathbf{x}+\tau_1) & \upsilon_1-a_l \end{pmatrix} \succeq 0, \\
& \forall\, l=1,...,L, \\
& \tau_k\geq 0,\ \forall k=1,...,r, \\
& \gamma_k\geq(\boldsymbol{\beta}_l+m_k\mathbf{D}_l)^T\mathbf{x}+a_lm_k^2+b_lm_k+c_l,\ \forall\, l=1,...,L, \\
& \forall\, k=r+1,...,K.
\end{aligned}
\tag{2.7}
$$

With or without probability ambiguity, each LMI can be further simplified into a second order cone constraint. If $\Omega=\mathbb{R}^+$ or some other subset of $\mathbb{R}$, the formulation is analogous. If there is no probability ambiguity, then the solution to (2.2) is given by the minimizer of (2.7), which is an SDP. With probability ambiguity and $\phi\in\Phi$, it is an SCP.

2.3 Polytopes and Unbounded Polytopes

Now we consider the general multivariate case for (2.4). Polytopes would be a natural regional structure to consider in order to efficiently fragment the data and capture its modality. However, it is difficult to reformulate nonnegativity of a quadratic function over a polytope as a linear matrix inequality. Generally, these constraints form less tractable copositive programs. Murty and Kabadi (1987) show that checking whether a matrix is copositive or not is co-NP-complete.

Suppose we fragment the support of the data into a set of polytopes $\{\Omega_k\}$, some of which may be unbounded. For each k, we construct a $(p+1)$-dimensional polytope $\Omega_k^{'}$ by adding another dimension and restricting its value to 1. We let the vertices and rays of $\Omega_k^{'}$ be $\mathbf{v}_{kv}$ for $v = 1, ..., V$, and $\mathbf{r}_{ku}$ for $u = 1, ..., U$, respectively.

Theorem 2. *Suppose that for each k, Ω_k is either a p-dimensional polytope with V vertices, unbounded polytope with V vertices and U rays, or point mass $\mathbf{m}_k$, and that Assumptions 1 and 2 hold. Then the solution to problem (2.2) is given by the minimizer of the following copositive program:*

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{\gamma_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \eta + \rho\lambda + \lambda \sum_{k=1}^{r} \hat{p}_k \phi^\star \left(\frac{z_k + {\mathbf{z}_k}^T \hat{\boldsymbol{\mu}}_k + \mathbf{Z}_k \circ (\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^T) - \eta}{\lambda} \right) \\
& + \lambda \sum_{k=r+1}^{K} \hat{p}_k \phi^\star \left(\frac{\gamma_k - \eta}{\lambda} \right), \\
\textit{subject to} \quad & \begin{pmatrix} \mathbf{v}_{k1}^T \\ \vdots \\ \mathbf{v}_{kV}^T \\ \mathbf{r}_{k1}^T \\ \vdots \\ \mathbf{r}_{kU}^T \end{pmatrix} \bar{\mathbf{Z}}_{kl} \begin{pmatrix} \mathbf{v}_{k1} & \cdots & \mathbf{v}_{kV} & \mathbf{r}_{k1} & \cdots & \mathbf{r}_{kU} \end{pmatrix} \in \mathcal{CO}(\mathbb{R}^+_{U+V}), \\
& \forall\, l = 1, ..., L, \forall k = 1, ..., r, \\
& z_k + {\mathbf{z}_k}^T \hat{\boldsymbol{\mu}}_k + \mathbf{Z}_k \circ (\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^T) - \eta \leq \left(\lim_{t\to\infty} \frac{\phi(t)}{t} \right) \lambda, \\
& \forall k = 1, ..., r, \\
& \gamma_k \geq (\boldsymbol{\beta}_l + \mathbf{D}_l \mathbf{m}_k)^T \mathbf{x} + \mathbf{m}_k^T \mathbf{A}_l \mathbf{m}_k + \mathbf{b}_l^T \mathbf{m}_k + c_l,\ \forall\, l = 1, ..., L, \\
& \forall k = r+1, ..., K, \\
& \gamma_k - \eta \leq \left(\lim_{t\to\infty} \frac{\phi(t)}{t} \right) \lambda,\ \forall k = r+1, ..., K, \\
& \lambda \geq 0.
\end{aligned}
\tag{2.8}
$$

where

$$\bar{\mathbf{Z}}_{kl} = \begin{pmatrix} z_k - c_l - \boldsymbol{\beta}_l^T \mathbf{x} & \frac{1}{2}(\mathbf{z}_k - \mathbf{b}_l - \mathbf{D}_l^T \mathbf{x})^T \\ \frac{1}{2}(\mathbf{z}_k - \mathbf{b}_l - \mathbf{D}_l^T \mathbf{x}) & \mathbf{Z}_k - \mathbf{A}_l \end{pmatrix}$$

and $\mathcal{CO}(\Omega) = \{\mathbf{A} \in \mathcal{S}_n | \mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0, \forall \mathbf{x} \in \Omega \subseteq \mathbb{R}^n\}$*, the generalized copositive cone over a set* Ω*. Consequently, if* $\rho = 0$*,* $p = 2$ *and each* Ω_k *is a triangle or quadrilateral, the problem is an SDP.*

Proof. For each k, we construct an n+1-dimensional polytope $\Omega_k^{'}$ by adding another dimension and restricting its value to 1. We let the vertices and rays of $\Omega_k^{'}$ be $\mathbf{v}_{ks}$ for $s = 1, ..., S$ and $\mathbf{r}_{kr}$ for $r = 1, ..., R$, respectively.
Let $\bar{\mathbf{Z}}_{kl} = \begin{pmatrix} z_k - c_l - \boldsymbol{\beta}_l^T \mathbf{x} & \frac{1}{2}(\mathbf{z}_k - \mathbf{b}_l - \mathbf{D}_l^T \mathbf{x})^T \\ \frac{1}{2}(\mathbf{z}_k - \mathbf{b}_l - \mathbf{D}_l^T \mathbf{x}) & \mathbf{Z}_k - \mathbf{A}_l \end{pmatrix}$. From Theorem 1, we have that the general form of the dual is as in (2.4). The kl infinite dimensional constraint $\begin{pmatrix} 1 & \boldsymbol{\xi}^T \end{pmatrix} \bar{\mathbf{Z}}_{kl} \begin{pmatrix} 1 \\ \boldsymbol{\xi} \end{pmatrix} \geq 0 \ \ \forall \boldsymbol{\xi} \in \Omega_k$, is equivalent to $\bar{\mathbf{Z}}_{kl} \in \mathcal{CO}(\Omega_k^{'})$. Let $\mathbf{a} = \sum_{s=1}^{S} \lambda_s \mathbf{v}_{ks} + \sum_{r=1}^{R} \gamma_r \mathbf{r}_{kr}$, where each $\lambda_s, \gamma_r \geq 0$, be a conic combination of the vertices and/or rays of $\Omega_k^{'}$. If $\mathbf{a}^T \bar{\mathbf{Z}}_{kl} \mathbf{a} \geq 0$ for any such $\mathbf{a}$, then this implies that $\bar{\mathbf{Z}}_{kl} \in \mathcal{CO}(\Omega_k^{'})$ since any $\mathbf{x} \in \Omega_k^{'}$ can be written as a convex combination of the vertices and conic combination of rays. Now suppose that $\bar{\mathbf{Z}}_{kl} \in \mathcal{CO}(\Omega_k^{'})$. By Lemma (4), this is equivalent to $\bar{\mathbf{Z}}_{kl}$ being copositive over the conic hull of $\Omega_k^{'}$. Now consider $\mathbf{a} = \sum_{s=1}^{S} \lambda_s \mathbf{v}_{ks} + \sum_{r=1}^{R} \gamma_r \mathbf{r}_{kr}$, and let $\mathbf{y} = \frac{\mathbf{a}}{\sum \lambda_s}$. Then $\mathbf{y}$ is a convex combination of the vertices and conic combination of rays, and thus $\mathbf{y} \in \Omega_k^{'}$. Therefore, $\mathbf{a}$ is in the conic hull of $\Omega_k^{'}$ and so $\mathbf{a}^T \bar{\mathbf{Z}}_{kl} \mathbf{a} \geq 0$. Hence, the constraint $\bar{\mathbf{Z}}_{kl} \in \mathcal{CO}(\Omega_k^{'})$ can be replaced with the equivalent condition that $\mathbf{a}^T \bar{\mathbf{Z}}_{kl} \mathbf{a} \geq 0$ for all $\mathbf{a} = \sum_{s=1}^{S} \lambda_s \mathbf{v}_{ks} + \sum_{r=1}^{R} \gamma_r \mathbf{r}_{kr}$, $\lambda_s, \gamma_r \geq 0$. This can be written as

$$\begin{pmatrix} \lambda_1 & \cdots & \lambda_S & \gamma_1 & \cdots & \gamma_R \end{pmatrix} \begin{pmatrix} \mathbf{v}_{k1}^T \\ \vdots \\ \mathbf{v}_{kS}^T \\ \mathbf{r}_{k1}^T \\ \vdots \\ \mathbf{r}_{kR}^T \end{pmatrix} \bar{\mathbf{Z}}_{kl} \begin{pmatrix} \mathbf{v}_{k1} & \cdots & \mathbf{v}_{kS} & \mathbf{r}_{k1} & \cdots & \mathbf{r}_{kR} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_S \\ \gamma_1 \\ \vdots \\ \gamma_R \end{pmatrix} \geq 0$$

for all $\boldsymbol{\lambda}, \boldsymbol{\gamma} \geq 0$ or equivalently that

$$\begin{pmatrix} \mathbf{v}_{k1}^T \\ \vdots \\ \mathbf{v}_{kS}^T \\ \mathbf{r}_{k1}^T \\ \vdots \\ \mathbf{r}_{kR}^T \end{pmatrix} \bar{\mathbf{Z}}_{kl} \begin{pmatrix} \mathbf{v}_{k1} & \cdots & \mathbf{v}_{kS} & \mathbf{r}_{k1} & \cdots & \mathbf{r}_{kR} \end{pmatrix} \in \mathcal{CO}(\mathbb{R}^+_{U+V}).$$

In the special case with no probability ambiguity ($\rho = 0$), we have the the general dual form is an in (2.5). Therefore, we have:

$$\begin{aligned} \min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{\gamma_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \sum_{k=1}^{r} \hat{p}_k \left[z_k + \mathbf{z}_k{}^T \hat{\boldsymbol{\mu}}_k + \mathbf{Z}_k \circ (\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^T) \right] + \sum_{k=r+1}^{r+m} \hat{p}_k \gamma_k, \\ \text{subject to} \quad & \begin{pmatrix} \mathbf{v}_{k1}^T \\ \cdots \\ \mathbf{v}_{kV}^T \\ \mathbf{r}_{k1}^T \\ \cdots \\ \mathbf{r}_{kU}^T \end{pmatrix} \bar{\mathbf{Z}}_{kl} \begin{pmatrix} \mathbf{v}_{k1} & \cdots & \mathbf{v}_{kV} & \mathbf{r}_{k1} & \cdots & \mathbf{r}_{kU} \end{pmatrix} \in \mathcal{CO}(\mathbb{R}^+_{U+V}), \\ & \forall\, l = 1, ..., L,\ \forall k = 1, ..., r, \\ & \gamma_k \geq \boldsymbol{\beta}_l^T \mathbf{x} + \mathbf{m}_k^T \mathbf{A}_l \mathbf{m}_k + \mathbf{b}_l^T \mathbf{m}_k + c_l, \forall\, l = 1, ..., L, \forall k = r+1, ..., K. \end{aligned} \tag{2.9}$$

If each polytope has fewer than or equal to four vertices and/or rays, each of the copositive constraints is an LMI by Lemma 5. If we are in $\mathbb{R}^2$ with quadrilaterals, or $\mathbb{R}^3$ with tetrahedrons, then this is the case. Thus, (2.2), with $p = 2$, $\rho = 0$ (and thus the objective function is linear as in (2.9)), and triangular and/or quadrilateral regions, is an SDP. □

In general, the number of vertices and rays $S + R$ of a given polytope may grow exponentially in the dimension p, and thus the size of the matrix in each constraint grows exponentially in p. For example, if we restrict our regions to hypercubes, then the number of vertices for any bounded region in the partition will be 2^p. We call on a result from Burer and Dong (2012) to show that we may have an alternative copositive

formulation where the size of the matrix grows polynomially. For example, for the hypercube case, the size of the matrix in each constraint grows as $4p + 2$. Thus, for larger problems, using this formulation will result in a smaller copositive program.

Lemma 1. *Under the conditions of Theorem 2, problem* (2.2) *may be formulated as a copositive program, where the size of the matrix constraints grows polynomially in p.*

Proof. The proof is given in Appendix B.2. □

Regardless of the formulation, this type of partition leads to a copositive program in the general case. We would like to design a partition that is reasonable and leads to more tractable formulations.

2.4 Ellipsoids and Differences of Ellipsoids

The key to making robust fragmentation applicable is to find a type of partition that is both meaningful and tractable. The difficulty in (2.4) lies in the infinite dimensional constraints, or non-negativity of a quadratic function over a domain Ω_k. In the univariate case, this is easily dealt with by use of the S-Lemma since an interval can be represented by a single quadratic inequality. In higher dimensions, the analogous region is an ellipsoid. But a partition of ellipsoids will leave gaps and may not be able to capture enough information about the shape of different modes in the distribution. To strengthen this approach, we extend the S-Lemma to be able to perform a similar procedure for the difference between ellipsoids.

Lemma 2. *Let f, g be quadratic functions on $\mathbb{R}^n$, with g strictly convex, and suppose there exists some $\mathbf{x}_0$ s.t. $g(\mathbf{x}_0) < b$. Then the following two statements are equivalent:*
(i) There is no $\mathbf{x}$ s.t.

$$f(\mathbf{x}) < 0, \quad a \leq g(\mathbf{x}) \leq b.$$

(ii) There exists τ_1, $\tau_2 \geq 0$, with $\tau_1\tau_2 = 0$, s.t.

$$f(\mathbf{x}) + \tau_1(a - g(\mathbf{x})) + \tau_2(g(\mathbf{x}) - b) \geq 0, \quad \forall \mathbf{x} \in \mathbb{R}^n.$$

Proof. Clearly, $(ii) \implies (i)$. From (i), we know that $f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ s.t. $g(\mathbf{x}) = b$. By assumption, we have that $g(\mathbf{x}_0) < b$, and by the strict convexity of g, we know

$\exists\ \mathbf{x}_1$ s.t. $g(\mathbf{x}_1) > b$. By Lemma 6, $\exists\ \tau$ s.t. $f(\mathbf{x}) + \tau(g(\mathbf{x}) - b) \geq 0\ \forall \mathbf{x}$. If $\tau \geq 0$, set $\tau_1 = 0$ and $\tau_2 = \tau$. If $\tau \leq 0$, we have that $f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ s.t. $g(\mathbf{x}) \geq b$, since $f(\mathbf{x}) \geq -\tau(g(\mathbf{x}) - b)$ for $-\tau \geq 0$. Therefore, since by assumption $f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ s.t. $a \leq g(\mathbf{x}) \leq b$, we know that $f(\mathbf{x}) \geq 0\ \forall g(\mathbf{x}) \geq a$. By Lemma 3, we have that $\exists\ \tau_1 \geq 0$ s.t. $f(\mathbf{x}) + \tau_1(a - g(\mathbf{x})) \geq 0$ for all $\mathbf{x}$. Then set $\tau_2 = 0$ and the proof is complete. □

Now suppose each subregion is an ellipsoid as in Hanasusanto et al. (2014) or a difference between ellipsoids. The parameters $\boldsymbol{\nu}_k \in \mathbb{R}^n$, $\boldsymbol{\Lambda}_k \in \mathcal{S}_n^+$, and $\delta_k \in \mathbb{R}$ determine the center, shape, and size of each ellipsoid, respectively.

Theorem 3. *Suppose that* $\Omega_k = \left\{\boldsymbol{\xi} \in \mathbb{R}^n\ :\ (\boldsymbol{\xi}^T - \boldsymbol{\nu}_k^T)\boldsymbol{\Lambda}_k^{-1}(\boldsymbol{\xi} - \boldsymbol{\nu}_k) \leq \delta_k^2\right\}\ \forall k \in E$, $\Omega_k = \left[\boldsymbol{\xi} \in \mathbb{R}^n\ :\ \delta_{kl}^2 \leq (\boldsymbol{\xi}^T - \boldsymbol{\nu}_k^T)\boldsymbol{\Lambda}_k^{-1}(\boldsymbol{\xi} - \boldsymbol{\nu}_k) \leq \delta_{ku}^2\right]\ \forall k \in D$, $\Omega_k = \mathbf{m}_k$ *with mass* p_k $\forall k \in S$, *and that Assumptions 1 and 2 hold. Then the solution to problem* (2.2), *is given by the minimizer of the following convex problem:*

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{\gamma_k,\tau_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \eta+\rho\lambda+\lambda\sum_{k\in E\cup D}\hat{p}_k\phi^{\star}\left(\frac{z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta}{\lambda}\right)\\
& +\lambda\sum_{k\in S}\hat{p}_k\phi^{\star}\left(\frac{\gamma_k-\eta}{\lambda}\right),\\
\text{subject to} \quad & \begin{pmatrix} z_k-c_l-\boldsymbol{\beta}_l^T\mathbf{x} & \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x})^T\\ \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x}) & \mathbf{Z}_k-\mathbf{A}_l\end{pmatrix}\\
& -\tau_k\begin{pmatrix}\delta_k^2-\boldsymbol{\nu}_k^T\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & (\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k)^T\\ \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & -\boldsymbol{\Lambda}_k^{-1}\end{pmatrix}\succeq 0,\ \forall\, l,\forall k\in E,\\
& \begin{pmatrix} z_k-c_l-\boldsymbol{\beta}_l^T\mathbf{x} & \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x})^T\\ \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x}) & \mathbf{Z}_k-\mathbf{A}_l\end{pmatrix}\\
& -\tau_{k_1}\begin{pmatrix}-\delta_{kl}^2+\boldsymbol{\nu}_k^T\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & -(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k)^T\\ -\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & \boldsymbol{\Lambda}_k^{-1}\end{pmatrix}\\
& -\tau_{k_2}\begin{pmatrix}\delta_{ku}^2-\boldsymbol{\nu}_k^T\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & (\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k)^T\\ \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & -\boldsymbol{\Lambda}_k^{-1}\end{pmatrix}\succeq 0,\ \forall\, l,\forall k\in D,\\
& \tau_k\geq 0,\ \forall k\in E,\\
& \tau_{k_1},\tau_{k_2}\geq 0,\ \forall k\in D,\\
& z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\\
& \forall k\in E\cup D,\\
& \gamma_k\geq(\boldsymbol{\beta}_l+\mathbf{D}_l\mathbf{m}_k)^T\mathbf{x}+\mathbf{m}_k^T\mathbf{A}_l\mathbf{m}_k+\mathbf{b}_l^T\mathbf{m}_k+c_l,\ \forall\, l=1,...,L,\\
& \forall k\in S,\\
& \gamma_k-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\quad\forall k\in S,\\
& \lambda\geq 0.
\end{aligned}
\tag{2.10}
$$

Proof. From Theorem 1, we have that the general form of the dual is as in (2.4).

Let $\bar{\mathbf{Z}}_{kl} = \begin{pmatrix} z_k - c_l - \boldsymbol{\beta}_l^T \mathbf{x} & \frac{1}{2}(\mathbf{z}_k - \mathbf{b}_l - \mathbf{D}_l^T \mathbf{x})^T \\ \frac{1}{2}(\mathbf{z}_k - \mathbf{b}_l - \mathbf{D}_l^T \mathbf{x}) & \mathbf{Z}_k - \mathbf{A}_l \end{pmatrix}$ as before. We need to reformulate the kl constraints $\begin{pmatrix} 1 & \boldsymbol{\xi}^T \end{pmatrix} \bar{\mathbf{Z}}_{kl} \begin{pmatrix} 1 \\ \boldsymbol{\xi} \end{pmatrix} \geq 0 \quad \forall \boldsymbol{\xi} \in \Omega_k$. First consider a subregion $\Omega_k = \left\{ \boldsymbol{\xi} \in \mathbb{R}^p \ : \ (\boldsymbol{\xi}^T - \boldsymbol{\nu}_k^T) \boldsymbol{\Lambda}_k^{-1} (\boldsymbol{\xi} - \boldsymbol{\nu}_k) \leq \delta_k^2 \right\}$ which is an ellipsoid. Since the ellipsoidal region can be described by a single quadratic inequality in $\boldsymbol{\xi}$, we can use S-Lemma to reformulate this as a single LMI as follows:

$$\left[\bar{\mathbf{Z}}_{kl} - \tau_k \begin{pmatrix} \delta_k^2 - \boldsymbol{\nu}_k^T \boldsymbol{\Lambda}_k^{-1} \boldsymbol{\nu}_k & (\boldsymbol{\Lambda}_k^{-1} \boldsymbol{\nu}_k)^T \\ \boldsymbol{\Lambda}_k^{-1} \boldsymbol{\nu}_k & -\boldsymbol{\Lambda}_k^{-1} \end{pmatrix} \right] \succeq 0, \ \tau_k \geq 0.$$

Now consider a subregion $\Omega_k = \left\{ \boldsymbol{\xi} \in \mathbb{R}^n \ : \ \delta_{kl}^2 \leq (\boldsymbol{\xi}^T - \boldsymbol{\nu}_k^T) \boldsymbol{\Lambda}_k^{-1} (\boldsymbol{\xi} - \boldsymbol{\nu}_k) \leq \delta_{ku}^2 \right\}$. In this case, the region Ω_k is defined by a set of two quadratic inequalities in $\boldsymbol{\xi}$. Using Lemma 2, it is clear that this is equivalent to the following LMI:

$$\bar{\mathbf{Z}}_{kl} - \tau_{k1} \begin{pmatrix} -\delta_{kl}^2 - \boldsymbol{\nu}_k^T \boldsymbol{\Lambda}_k^{-1} \boldsymbol{\nu}_k & (\boldsymbol{\Lambda}_k^{-1} \boldsymbol{\nu}_k)^T \\ \boldsymbol{\Lambda}_k^{-1} \boldsymbol{\nu}_k & -\boldsymbol{\Lambda}_k^{-1} \end{pmatrix} - \tau_{k2} \begin{pmatrix} \delta_{ku}^2 - \boldsymbol{\nu}_k^T \boldsymbol{\Lambda}_k^{-1} \boldsymbol{\nu}_k & (\boldsymbol{\Lambda}_k^{-1} \boldsymbol{\nu}_k)^T \\ \boldsymbol{\Lambda}_k^{-1} \boldsymbol{\nu}_k & -\boldsymbol{\Lambda}_k^{-1} \end{pmatrix} \succeq 0,$$
$$\text{with } \tau_{k1}, \tau_{k2} \geq 0.$$

We combine the different types of constraints to get the formulation in (2.10). □

In the general case, with probability ambiguity and $\phi \in \Phi$, we have an SCP. In the special case of no probability ambiguity, we have that the general form is an in (2.5).

Thus, we have an SDP with mr LMI constraints, where each matrix is of size $p+1$:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\lambda_k,\tau_k,\{z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \sum_{k\in E\cup D} \hat{p}_k \left[z_k + {\mathbf{z}_k}^T \hat{\boldsymbol{\mu}}_k + \mathbf{Z}_k \circ (\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^T) \right] + \sum_{k\in S} \hat{p}_k \lambda_k, \\
\text{subject to} \quad & \begin{pmatrix} z_k - c_l - \boldsymbol{\beta}_l^T \mathbf{x} & \frac{1}{2}(\mathbf{z}_k - \mathbf{b}_l)^T \\ \frac{1}{2}(\mathbf{z}_k - b_l) & \mathbf{Z}_k - \mathbf{A}_l \end{pmatrix} - \tau_k \begin{pmatrix} \delta_k^2 - \boldsymbol{\nu}_k^T \boldsymbol{\Lambda}_k^{-1} \boldsymbol{\nu}_k & (\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k)^T \\ \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & -\boldsymbol{\Lambda}_k^{-1} \end{pmatrix} \\
& \succeq 0, \ \forall\, l, \forall k \in E, \\
& \begin{pmatrix} z_k - c_l - \boldsymbol{\beta}_l^T \mathbf{x} & \frac{1}{2}(\mathbf{z}_k - \mathbf{b}_l)^T \\ \frac{1}{2}(\mathbf{z}_k - b_l) & \mathbf{Z}_k - \mathbf{A}_l \end{pmatrix} \\
& - \tau_{k1} \begin{pmatrix} -\delta_{kl}^2 + \boldsymbol{\nu}_k^T \boldsymbol{\Lambda}_k^{-1} \boldsymbol{\nu}_k & -(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k)^T \\ -\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & \boldsymbol{\Lambda}_k^{-1} \end{pmatrix} \\
& - \tau_{k2} \begin{pmatrix} \delta_{ku}^2 - \boldsymbol{\nu}_k^T \boldsymbol{\Lambda}_k^{-1} \boldsymbol{\nu}_k & (\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k)^T \\ \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & -\boldsymbol{\Lambda}_k^{-1} \end{pmatrix} \succeq 0, \ \forall\, l, \forall k \in D, \\
& \tau_k \geq 0, \ \forall k \in E, \\
& \tau_{k1}, \tau_{k2} \geq 0, \ \forall k \in D, \\
& \lambda_k \geq \boldsymbol{\beta}_l^T \mathbf{x} + \mathbf{m}_k^T \mathbf{A}_l \mathbf{m}_k + \mathbf{b}_l^T \mathbf{m}_k + c_l, \ \forall\, l = 1, ..., L, \\
& \forall k \in S.
\end{aligned}
\tag{2.11}
$$

This provides us with a tractable formulation for a flexible type of fragmentation. Ellipsoids and differences of ellipsoids can capture modes of the distribution and, within each mode, the structure in layers. To see the intuition behind such a scheme, consider a normal distribution. Ellipsoids are level sets of the multivariate density function. Regions that are differences between ellipsoids, centered at the mean and shaped according to the covariance, represent layers of level sets, so that each point in a given region has similar density. This will be similar for many unimodal distributions. Fragmenting in such a way can be used to capture information about skewness and modality in layers.

Combining ellipsoids and differences of ellipsoids to form a partition allows us to capture as much detail as we choose. Figure 2.5 provides an example of a data sample for a two-dimensional problem and a potential fragmentation. Each mode is captured by an ellipsoidal region. Within each mode, as the size and the number of points increase,

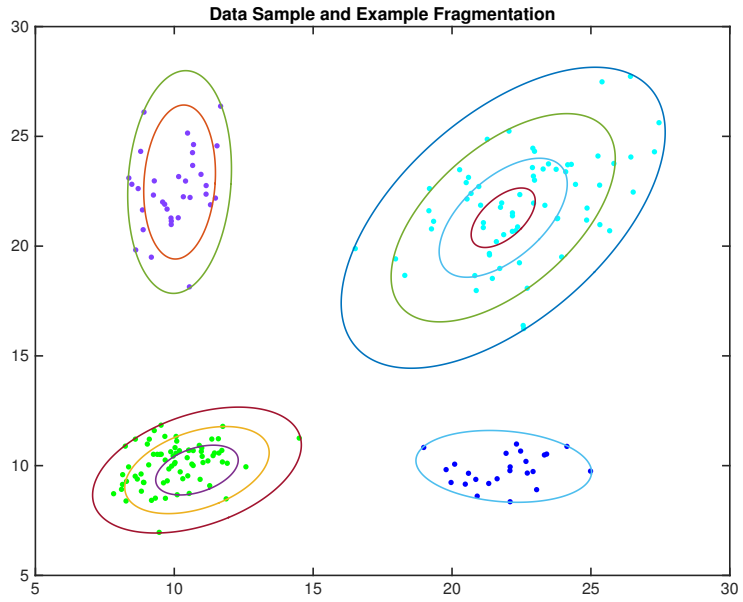

Figure 2.5: Two-Dimensional Data Sample and Example Fragmentation

it is further broken up into differences of ellipsoids. In Chapters 4 and 5, we discuss various approaches to constructing a fragmentation, given a data sample.

2.5 Convergence

Consider (2.2) in the case where there is no probability ambiguity ($\rho = 0$). We wish to show convergence of the optimal value and solution set to the same problem with the true moments p_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ as $N \to \infty$. Thus, as sample size increases, we can guarantee convergence of the problem to a problem where the ambiguity set contains the true distribution. Given the true moments, let $v(\mathbf{x})$ be the optimal value to the inner maximization problem in (2.2) for a given $\mathbf{x}$. Let $\mathbf{X} = \text{argmin}_{\mathbf{x}} v(\mathbf{x})$ and $\Upsilon = \min_{\mathbf{x}} v(\mathbf{x})$ be the optimal solution and optimal value of (2.2), respectively.

Now suppose we solve (2.2) with estimated moments p_k^N, $\boldsymbol{\mu}_k^N$, and $\boldsymbol{\Sigma}_k^N$, for $k = 1, ..., r$, and sample size N. Let $v_N(\mathbf{x})$, $\mathbf{X}_N$, and Υ_N be the corresponding objective function, optimal solution and optimal value for the outer minimization problem. If we

consider a sequence of such problems with different sample sizes N, we would like to show that as $N \to \infty$, the optimal values $\Upsilon_N \to \Upsilon$, and the optimal solutions $X_N \to X$. That is, the solution of the RF with estimated moments will converge to that of the RF with the true moments.

Proposition 2. *If we use the sample values as estimates,* $p_k^N = \frac{N_k}{N}$ *(where* $N_k = \sum_{i=1}^N \mathbb{1}(\boldsymbol{\xi}_i \in \Omega_k)$*),* $p_k^N \boldsymbol{\mu}_k^N = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\xi}_i \mathbb{1}(\boldsymbol{\xi}_i \in \Omega_k)$*, and* $p_k^N(\boldsymbol{\Sigma}_k^N + \boldsymbol{\mu_k}^N{}^T \boldsymbol{\mu}_k^N) = \frac{1}{N-1}\sum_{i=1}^N \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T \mathbb{1}(\boldsymbol{\xi}_i \in \Omega_k)$*, then the following hold (under certain conditions outlined in the appendix):*

- $\lim_{N\to\infty} \sup_{\mathbf{x}\in\mathcal{X}} |v_N(\mathbf{x}) - v(\mathbf{x})| = 0$ *almost surely.*
- $\lim_{N\to\infty} \mathbf{X}_N \subseteq \mathbf{X}$ *and* $\lim_{N\to\infty} \Upsilon_N = \Upsilon$ *almost surely.*

Proof. We assume conditions as in Assumptions 1, 2, and 3(a) of Sun and Xu (2015), which relate to the cost function f, feasible set $\mathcal{X}$, and the character of the ambiguity sets Θ_N and Θ. Let $\hat{\Theta}$ be as defined in Assumption 2 where $\Theta_N, \Theta \subseteq \hat{\Theta}$. We list only the parts of Assumption 1 here:

1. For each $\boldsymbol{\xi} \in \Xi$, $f(\cdot, \boldsymbol{\xi})$ is Lipschitz continuous with modulus bounded by $\kappa(\boldsymbol{\xi})$, where $\sup_{\pi\in\hat{\Theta}} \mathbb{E}_\pi[\kappa(\boldsymbol{\xi})] < \infty$.
2. There exists $\mathbf{x}_0 \in \mathcal{X}$ such that $\sup_{\pi\in\hat{\Theta}} \|\mathbb{E}_P[f(\mathbf{x}_0, \boldsymbol{\xi})]\| < \infty$.
3. $\mathcal{X}$ is compact.

Assumption 3(a) in Sun and Xu (2015) is convergence in pseudometric of the ambiguity sets Θ_N to Θ. We aim to show this convergence is satisfied by using the sample values as estimates.

We rewrite the constraints in (2.2) as $\int_{\Omega_k} 1 d\pi(\tilde{\boldsymbol{\xi}}) = p_k$, $\int_{\Omega_k} \tilde{\boldsymbol{\xi}} d\pi(\tilde{\boldsymbol{\xi}}) = p_k \boldsymbol{\mu}_k$, and $\int_{\Omega_k} \tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}^T d\pi(\tilde{\boldsymbol{\xi}}) = p_k(\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T)$ for $k = 1, ..., r$. We can equivalently define our ambiguity sets as follows: $\Theta_N := \{\pi \in \mathscr{P} \ : \ \mathbb{E}_\pi[\psi_q(\boldsymbol{\xi})] = \nu_q^N, \text{ for } q = 1, ..., Q\}$ where $\psi_q(\boldsymbol{\xi}) = \mathbb{1}(\boldsymbol{\xi} \in \Omega_k)$, $\psi_q(\boldsymbol{\xi}) = \xi_j \mathbb{1}(\boldsymbol{\xi} \in \Omega_k)$, or $\psi_q(\boldsymbol{\xi}) = \xi_{j_1}\xi_{j_2}\mathbb{1}(\boldsymbol{\xi} \in \Omega_k)$ for some $j, j_1, j_2 \in (1, ..., p)$ and $k \in (1, ..., r)$, and $\nu_q^N = p_k^N$, $p_k^N \mu_{kj}^N$, or $p_k^N(\Sigma_{kj_1j_2}^N + \mu_{kj_1}^N \mu_{kj_2}^N)$, respectively. The true ambiguity set Θ is defined similarly but with true moments p_k, $p_k\mu_{kj}$, and $p_k(\Sigma_{kj_1j_2} + \mu_{kj_1}\mu_{kj_2})$.

We assume the following regularity condition: $(1, \boldsymbol{\nu})) \in \text{int}[(\langle \pi, 1 \rangle, \langle \pi, \psi(\boldsymbol{\xi}) \rangle) - \mathbf{0} : \pi \in \mathscr{P}]$, which amounts to requiring that the moments be in the interior of the moments cone. We need to show that $\nu_q^N \to \nu_q$ as $N \to \infty$ for any $q = 1, ..., Q$. This is equivalent to showing that $p_k^N \to p_k$, $p_k^N \boldsymbol{\mu}_k^N \to p_k \boldsymbol{\mu}_k$, and $p_k^N (\boldsymbol{\Sigma}_k^N + \boldsymbol{\mu}_k^N \boldsymbol{\mu}_{\boldsymbol{k}}^{\boldsymbol{N}^T}) \to p_k (\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_{\boldsymbol{k}}^T)$, $\forall k = 1, ..., r$. Since p_k^N is equivalent to a binomial random variable divided by the number of trials N with success probability p_k, it converges to p_k almost surely by the strong law of large numbers $\forall k = 1, ..., r$. Similarly, $p_k^N \boldsymbol{\mu}_k^N$ is the sample average of a random variable that is a product of a Bernoulli r.v. $\tilde{u}_k$ with success probability p_k and a random variable $\tilde{\boldsymbol{\xi}}$. We have that this converges almost surely to $\mathbb{E}[\tilde{u}_k \tilde{\boldsymbol{\xi}}] = \mathbb{E}[\tilde{u}_k \tilde{\boldsymbol{\xi}} | u_k = 1] P(\tilde{u}_k = 1) + \mathbb{E}[\tilde{u}_k \tilde{\boldsymbol{\xi}} | u_k = 0] P(\tilde{u}_k = 0) = \boldsymbol{\mu}_k p_k \ \forall k = 1, ..., r$. The convergence of the second moment constraint follows similarly, where we have the product of a Bernoulli r.v. and random variable $\tilde{\boldsymbol{\xi}} \tilde{\boldsymbol{\xi}}^T$. Therefore, by Proposition 4 of Sun and Xu (2015), we have that Assumption 3(a) holds and therefore the two convergence results hold by Theorems 1 and 2 of Sun and Xu (2015). □

The consistency of the estimators of the conditional moments translates to consistency of the optimal value and solution to the robust problem.

Proposition 2 demonstrates convergence of the RF in (2.2) with estimated moments to the RF in (2.2) with true moments for a fixed fragmentation as $N \to \infty$. That is, $\mathbf{X}_N \xrightarrow{N \to \infty} \mathbf{X}$ and $\Upsilon_N \xrightarrow{N \to \infty} \Upsilon$. What about convergence of (2.2) to the solution and optimal value of (1), $\mathbf{X}^\star$ and $\Upsilon^\star$, under the true distribution $\pi^\star$? Suppose we have the true moments. As the number of regions $r \to \infty$, the diameter d of each region approaches zero (if we partition Ω into evenly sized hypercubes). As the regions collapse to singletons, the ambiguity set converges to a single distribution, the true distribution. Thus, $\mathbf{X} \xrightarrow{r \to \infty} \mathbf{X}^\star$ and $\Upsilon \xrightarrow{r \to \infty} \Upsilon^\star$. As $r, N \to \infty$ together, convergence of the optimal value $\Upsilon_N \xrightarrow{r, N \to \infty} \Upsilon^\star$ and solution $\mathbf{X}_N \xrightarrow{r, N \to \infty} \mathbf{X}^\star$ will depend on the interplay of r and N. In Section 4.1, we will use address convergence of the optimal value Υ_N in Theorem 5, and use this result to develop a heuristic for determining the number of regions based on the sample size.

In practice, we must have finite N and will therefore have estimation errors in the moments. From (2.5), we can see that without probability ambiguity, the objective function is linear in the moments, which do not affect the constraint set. Thus, small perturbations in the moments parameters may only affect small perturbations in the

optimal value. With probability ambiguity, the effect will be nonlinear due to $\phi^\star$.

Chapter 3

Overlapping Regions

In the previous derivations, we have assumed that the regions Ω_k are disjoint. If we instead allow some of the regions to be overlapping, the problem becomes considerably more challenging. We cannot separate the objective function since the regions are no longer independent and the expectation is no longer the weighted sum of the conditional expectations. This scenario is important to consider, since it may be challenging or undesirable to cover the sample data with nonoverlapping ellipsoids and differences of ellipsoids. We do not want to restrict our implementation by requiring a nonoverlapping fragmentation. In this chapter, we will demonstrate the additional challenges presented by an overlapping fragmentation and introduce one way of solving for an upper bound. However, instead of solving for this upper bound, we will alter our computation of conditional moments and solve the problem as if there was no overlap, as we will detail in the second section.

3.1 Upper Bound

In 2.2, we assume that the regions are disjoint. If we instead allow some of the regions to be overlapping, we cannot separate the objective function since the regions are no longer independent. Therefore, we cannot formulate the outer layer dual problem as in Theorem 1.

We can always construct a partition of the support of our fragmentation, where each region Ω_j, $j = 1, ..., R$ $(\geq r)$, consists of a unique intersection of original regions Ω_k.

That is, in any part of the domain where multiple regions are overlapping, we simply define a new region Ω_j that covers this intersection. Each region Ω_j will be a subset of some set of original regions I_j and each original region $\Omega_k = \{\cup\Omega_j : \ k \in I_j\}$. If we were able to break up the expectation over the Ω_j, then the terms would be independent. It would be reasonable to assume that we would have conditional moments estimates on these regions, as we could easily compute them from historical data. However, even if the original regions Ω_k are "nice", in that they admit tractable formulations, there are no guarantees on the intersecting regions Ω_j. For example, we know that ellipsoidal regions lead to tractable formulations. A nonnegativity constraint over an ellipsoid can be handled due to S-Lemma as described in the proof of (2.10), but the intersection of two or more ellipsoids is no longer an ellipsoid and would be intractable. Thus, breaking it up in such a way would be unproductive. We would arrive at a dual problem identical to the one derived in Theorem 1, broken up for each region j:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{\gamma_j,z_j,\mathbf{z}_j,\mathbf{Z}_j\}} \quad & \eta + \rho\lambda + \lambda\sum_{j=1}^{R}\hat{p}_j\phi^\star\left(\frac{z_j + {\mathbf{z}_j}^T\hat{\boldsymbol{\mu}}_j + \mathbf{Z}_j\circ(\hat{\boldsymbol{\Sigma}}_j + \hat{\boldsymbol{\mu}}_j\hat{\boldsymbol{\mu}}_j^T) - \eta}{\lambda}\right) \\
& + \lambda\sum_{j=R+1}^{K}\hat{p}_j\phi^\star\left(\frac{\gamma_j - \eta}{\lambda}\right), \\
\text{subject to} \quad & \boldsymbol{\xi}^T\mathbf{Z}_j\boldsymbol{\xi} + \mathbf{z}_j^T\boldsymbol{\xi} + z_j \geq f(\mathbf{x},\boldsymbol{\xi}),\ \forall\boldsymbol{\xi}\in\Omega_j,\ \forall j = 1,...,R, \\
& z_j + {\mathbf{z}_j}^T\hat{\boldsymbol{\mu}}_j + \mathbf{Z}_j\circ(\hat{\boldsymbol{\Sigma}}_j + \hat{\boldsymbol{\mu}}_j\hat{\boldsymbol{\mu}}_j^T) - \eta \leq \left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\ \forall j = 1,...,R, \\
& \gamma_j \geq (\boldsymbol{\beta}_l + \mathbf{D}_l\mathbf{m}_j)^T\mathbf{x} + \mathbf{m}_j^T\mathbf{A}_l\mathbf{m}_j + \mathbf{b}_l^T\mathbf{m}_j + c_l,\ \ \forall\, l = 1,...,L, \\
& \forall j = R+1,...,K, \\
& \gamma_j - \eta \leq \left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\ \forall j = R+1,...,K, \\
& \lambda \geq 0.
\end{aligned}
\tag{3.1}
$$

For each j, we have the infinite dimensional constraint $\boldsymbol{\xi}^T\mathbf{Z}_j\boldsymbol{\xi} + \mathbf{z}_j^T\boldsymbol{\xi} + z_j \geq f(\mathbf{x},\boldsymbol{\xi}),\ \forall\boldsymbol{\xi} \in \Omega_j$. Unless Ω_j is an ellipsoid or a polytope, we cannot reformulate this tractably. There is no type of fragmentation where we can guarantee the intersections will be ellipsoids. We could fragment into polytopal regions and the intersections will also be polytopes,

so the problem does not arise. However, polytopal regions are already NP hard and the number of edges in intersecting regions could be much more than for the original regions, making the problem size even larger.

However, if we treat the problem as if the regions were disjoint, then we can get an upper bound.

Theorem 4. *An upper bound to Problem* (2.3)*, when regions are overlapping and when Assumptions 1 and 2 hold, is given by the minimizer of the following optimization problem:*

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{\gamma_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \alpha\eta+\rho\lambda+\lambda\sum_{k=1}^{r}\hat{p}_k\phi^\star\left(\frac{z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta}{\lambda}\right)\\
& +\lambda\sum_{k=r+1}^{K}\hat{p}_k\phi^\star\left(\frac{\gamma_k-\eta}{\lambda}\right),\\
\textit{subject to}\quad & \begin{pmatrix}1 & \boldsymbol{\xi}^T\end{pmatrix}\begin{pmatrix}z_k-c_l-\boldsymbol{\beta}_l^T\mathbf{x} & \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x})^T\\ \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x}) & \mathbf{Z}_k-\mathbf{A}_l\end{pmatrix}\begin{pmatrix}1\\ \boldsymbol{\xi}\end{pmatrix}\geq 0\\
& \forall\boldsymbol{\xi}\in\Omega_k,\forall l=1,...,L,\ \forall k=1,...,r,\\
& z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\\
& \forall k=1,...,r,\\
& \gamma_k\geq(\boldsymbol{\beta}_l+\mathbf{D}_l\mathbf{m}_k)^T\mathbf{x}+\mathbf{m}_k^T\mathbf{A}_l\mathbf{m}_k+\mathbf{b}_l^T\mathbf{m}_k+c_l,\ \ \forall\, l=1,...,L,\\
& \forall k=r+1,...,K,\\
& \gamma_k-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\quad \forall k=r+1,...,K,\\
& \lambda\geq 0.
\end{aligned}
\tag{3.2}
$$

Proof. We treat the problem as if the regions were disjoint and break up the expectation

as before:

$$\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X}} \quad & \sup_{\mathbf{p}\in\mathscr{P}} \sum_{k=1}^{r} p_k \sup_{\pi_k\in\Theta_k} \int_{\Omega_k} f(\mathbf{x},\boldsymbol{\xi}) d\pi_k(\boldsymbol{\xi}) + \sum_{k=r+1}^{K} p_k f(\mathbf{x},\mathbf{m}_k), \\
\text{subject to} \quad & \mathbf{e}^T\mathbf{p} = \alpha, \\
& \mathbf{p} \geq 0, \\
& \sum_{k=1}^{K} \hat{p}_k \phi(\frac{p_k}{\hat{p}_k}) \leq \rho, \\
& \int_{\Omega_k} d\pi_k(\boldsymbol{\xi}) = 1, \\
& \int_{\Omega_k} \boldsymbol{\xi} d\pi_k(\boldsymbol{\xi}) = \hat{\boldsymbol{\mu}}_k, \\
& \int_{\Omega_k} \boldsymbol{\xi}\boldsymbol{\xi}^T d\pi_k(\boldsymbol{\xi}) = \hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T, \ \forall k = 1,...,r,
\end{aligned} \tag{3.3}$$

where π_k is the conditional distribution for $\boldsymbol{\xi}$ given that it is in Ω_k, and the distribution for each component is assumed to be independent, even if they are overlapping. Since the regions have overlap, we have that $\mathbf{e}^T\mathbf{p} = \alpha \geq 1$. This is clearly an upper bound to the RF problem (with regional moments matching the computed moments and computed probability vector with ambiguity set), since the constraints are the same, and the objective has been replaced with an expectation where intersections are counted multiple times, once for each region k they belong to.

For the dual of each inner problem, we ignore the moments constraints on any other regions that overlap, which amounts to a restriction of the dual variables. To see that this is a restriction of the dual variables, consider if we did not ignore overlapping moments constraints. In the original objective, for each k, we would break $\int_{\Omega_k} f(\mathbf{x},\boldsymbol{\xi}) d\pi_k(\boldsymbol{\xi}) = \cup_{j:k\in I_j} \int_{\Omega_j} f(\mathbf{x},\boldsymbol{\xi}) d\pi_k(\boldsymbol{\xi})$. Thus, for constraints, we would have $\sum_{k\in I_j} \boldsymbol{\xi}^T\mathbf{Z}_k\boldsymbol{\xi} + \mathbf{z}_k^T\boldsymbol{\xi} + z_k \geq f(\mathbf{x},\boldsymbol{\xi}),\ \forall\boldsymbol{\xi}\in\Omega_j,\ \forall j : k \in I_j$. Any set of dual variables that satisfy the constraints $\boldsymbol{\xi}^T\mathbf{Z}_k\boldsymbol{\xi} + \mathbf{z}_k^T\boldsymbol{\xi} + z_k \geq f(\mathbf{x},\boldsymbol{\xi}),\ \forall\boldsymbol{\xi}\in\Omega_k$ for all $k = 1,...,r$ must satisfy the true constraints. This is clear, since for any j, each kth individual term in the sum on the left hand side will be greater than the right hand side. Thus, since we have replaced the original objective function with an upper bound and then restricted the set of dual variables, where the dual problem is a minimization, (3.2) will be an

upper bound on the worst case expectation, given an RF problem such as in (2.3) with moments and probability constraints from overlapping regions. □

The gap between the worst case expected cost and the upper bound could be quite large in this instance, with probabilities of overlapping regions adding up to more than 1. The upper bound is useful because it is tractable. Even if this upper bound is not tight, if the gap for each $\mathbf{x}$ is similar, monotonic, or otherwise simple, then the optimal solution will not change much. We wished to test this. In overlapping cases, we have no way of solving for the true RF solution, so we cannot compare bounds or solutions directly, but we can compare to the true RF solution for a slightly different fragmentation, which covers only some of the data, but where there is no overlap. In addition, we are able to compare performance of the upper bound RF solution relative to MDRO and SAA, for a simple case with two regions and then cross-compare with relative performance of the true RF solution for similar regions that are not overlapping in order to get a benchmark.

The next few simulations are based on the two dimensional newsvendor problem, which will be described in Chapter 6. In each case, we use the training data to compute conditional probabilities and moments and use the estimates to solve the optimization problem for an optimal inventory $\mathbf{x}$. We then evaluate the cost function f based on the inventory and realizations of the demand $\boldsymbol{\xi}$ from a testing sample. In Figure 3.1, we show a training sample and two sets of regions, one a fragmentation that does not cover the training sample and one that covers the points but overlaps. On the right, in Figure 3.2, we compare the out-of-sample costs from employing the true RF solution ("RF overlap") over the smaller regions (where one will lose information from points outside the regions) to the costs from the upper bound solution over the overlapping regions ("RF Overlap"). It is clear that the upper bound performs better (lower costs over 500 simulations at each set of parameter values). This informs us that it is better to include all points in the regions even if it requires overlap, rather than sacrifice some to obtain no overlap. We also compare the performance to MDRO and SAA. The upper bound based method performs better than MDRO, even without solving the original problem, and about as well as SAA.

In Figure 3.3, we have generated data that is similar to the previous sample, but allows for a nonoverlapping fragmentation with two ellipsoids. We want to compare the

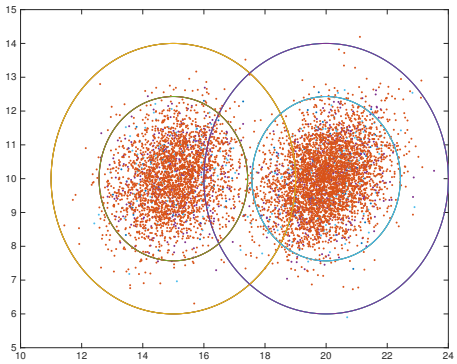

Figure 3.1: Overlapping Data and Fragmentations

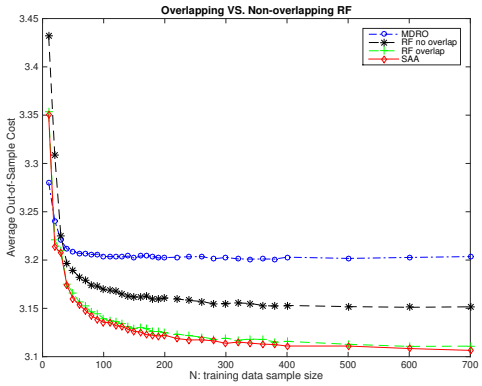

Figure 3.2: Overlapping VS. Non-overlapping Costs for RF (w/ MDRO, SAA)

relative performance of RF to MDRO and SAA in this example, given in Figure 3.4, to that in 3.2 in order to get a benchmark for how RF should perform relative to the other methods and whether or not overlap has an adverse effect. In 3.4, we see that RF performs about as well as SAA in both cases and substantially better than MDRO. This indicates that the relative performance of RF in Figure 3.2 has not deteriorated, even though we are only solving for the upper bound for each $\mathbf{x}$ rather than the true worst case expected cost. Since we are solving for a worst case for each $\mathbf{x}$, and the ambiguity sets are still the same across all $\mathbf{x}$ values for each problem, it seems that solving for the upper bound in the case of overlapping regions does not have a large effect on the solution, at least in terms of its practical performance.

3.2 Adaptive Estimation

Although there may be justification numerically for solving for an upper bound as illustrated in the previous section, there is a better alternative. Instead of taking this approach, we try to approximate the expectation, rather than get an upper bound, by adjusting how we compute the parameters.

When computing conditional probabilities and moments for Ω_k, we weight the sample points according to the probability that they are part of mode k. We denote the kth

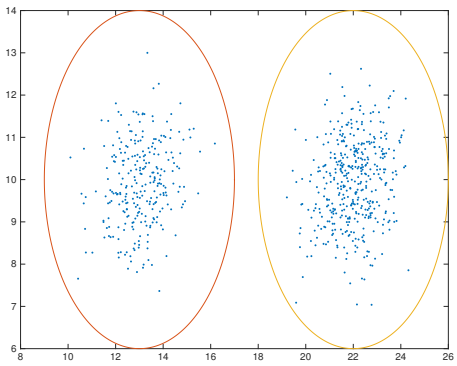

Figure 3.3: Non-Overlapping Data

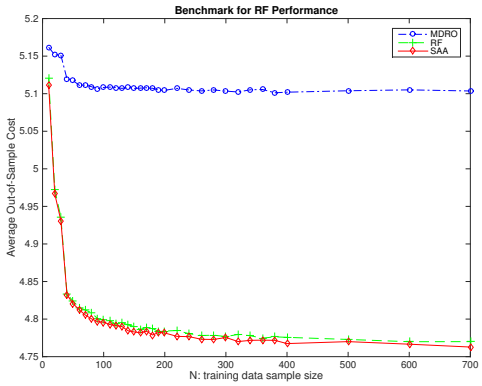

Figure 3.4: UB vs. Restricted

mode of the sample by $M_k \subseteq \Omega_k$. Some points in Ω_k may not belong to M_k if they are in an overlapping region. If a sample point belongs to multiple regions, then its contribution to parameter estimates is split amongst the regions according to these probabilities. For $\boldsymbol{\xi} \in \Omega_j$, we may estimate the conditional probability $w_{kj} = \mathbb{P}(\boldsymbol{\xi} \in M_k | \boldsymbol{\xi} \in \Omega_j)$ for all $k \in I_j$. For a nonoverlapping region j, this is equal to 1 and $\Omega_j = M_k$ for $k = I_j$. For an overlapping region with $|I_j| > 1$, we must have that $\sum_{k \in I_j} w_{kj} = 1, \ \forall j$ for the following consistency calculations to hold.

In principle any sets of weights may be chosen that satisfy these conditions. Since we interpret them as conditional probabilities, we weight proportionally to the total number of sample points in each region. Formally, for each k, we compute the solo probability $\tilde{p}_k = \frac{1}{N} \sum_{i=1}^{N} I(\boldsymbol{\xi}_i \in \Omega_k)$. Then w_{kj} is the conditional probability of belonging to mode k in region j, given by $w_{kj} = \frac{\tilde{p}_k}{\sum_{k' \in I_j} \tilde{p}_{k'}}$. We compute adjusted conditional statistics as follows:

$$
\begin{aligned}
\hat{p}_k &= \frac{1}{N} \sum_{i=1}^{N} \sum_{j:k \in I_j} w_{kj} I(\boldsymbol{\xi}_i \in \Omega_j), \\
\hat{\boldsymbol{\mu}}_k &= \frac{1}{N p_k} \sum_{i=1}^{N} \sum_{j:k \in I_j} w_{kj} \boldsymbol{\xi}_i I(\boldsymbol{\xi}_i \in \Omega_j), \\
\hat{\boldsymbol{\Sigma}}_k &= \frac{1}{N p_k} \sum_{i=1}^{N} \sum_{j:k \in I_j} w_{kj} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T I(\boldsymbol{\xi}_i \in \Omega_j) - \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^T,
\end{aligned}
\tag{3.4}
$$

so that all points in an intersection Ω_j contribute to all regions making up the intersection according to the weights w_{kj}.

Proposition 3. *The adjusted regional probability and conditional moments estimates as defined in* (3.4) *are consistent as probability estimates and with the overall estimates of mean and covariance from a historical sample.*

Proof. Recall that $R \geq r$ is the number of unique intersection regions j. We have that $\sum_{k=1}^{K} \hat{p}_k = \sum_{k=1}^{K} \frac{1}{N} \sum_{i=1}^{N} \sum_{j:k\in I_j} w_{kj} I(\boldsymbol{\xi}_i \in \Omega_j) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{R} \sum_{k\in I_j} w_{kj} I(\boldsymbol{\xi}_i \in \Omega_j) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{R} I(\boldsymbol{\xi}_i \in \Omega_j) = 1$ since $\cup \Omega_j = \Omega$, and so the component probabilities sum to 1. We can see that $\sum_{k=1}^{r} \hat{p}_k \hat{\boldsymbol{\mu}}_k = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{R} \sum_{k\in I_j} w_{kj} \boldsymbol{\xi}_i I(\boldsymbol{\xi}_i \in \Omega_j) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{R} \boldsymbol{\xi}_i I(\boldsymbol{\xi}_i \in \Omega_j) = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\xi}_i = \hat{\boldsymbol{\mu}}$, and similarly, that $\sum_{k=1}^{r} \hat{p}_k \hat{\boldsymbol{\Sigma}}_k = \hat{\boldsymbol{\Sigma}}$. Thus, the adjusted probabilities and conditional means and covariances are consistent with the overall estimates of mean and covariance from a historical sample. □

With these new estimates, we break up the worst case expectation into the sum of weighted conditional expectations over the Ω_k:

$$\sup_{\mathbf{p}\in\mathscr{P}} \sum_{k\in C} \tilde{p}_k \sup_{\pi_k\in\Theta_k} \int_{\Omega_k} f(\mathbf{x}, \boldsymbol{\xi}) d\pi_k(\boldsymbol{\xi}) + \sum_{k\in S} \tilde{p}_k f(\mathbf{x}, \mathbf{m}_k),$$

where the constraints on the probability vector $\mathbf{p}$ are as in Theorem 1, π_k are the conditional distributions, and the sets C and S are defined as before. Now, for each conditional distribution we have first and second order moments constraints corresponding to adapted estimates $\tilde{\boldsymbol{\mu}}_k$ and $\tilde{\boldsymbol{\Sigma}}_k$. For points in the overlapping areas, we are splitting up their contribution into each subregion in which they are contained. This is an approximation to the worst case expectation. In general, there should be minimal overlapping, which can be controlled by the user implementation. Additionally, the parameter estimates are universal over $\mathbf{x}$, so as not to bias the solution. We formulate the dual of each subproblem and combine as if they were independent (ignoring overlap), giving us

the following in replacement of (2.4):

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{\gamma_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \eta+\rho\lambda+\lambda\sum_{k=1}^{r}\tilde{p}_k\phi^\star\left(\frac{z_k+{\mathbf{z}_k}^T\tilde{\boldsymbol{\mu}}_k+\mathbf{Z}_k\cdot(\tilde{\boldsymbol{\Sigma}}_k+\tilde{\boldsymbol{\mu}}_k\tilde{\boldsymbol{\mu}}_k^T)-\eta}{\lambda}\right)\\
& +\lambda\sum_{k=r+1}^{K}\tilde{p}_k\phi^\star\left(\frac{\gamma_k-\eta}{\lambda}\right),\\
\text{subject to} \quad & \begin{pmatrix}1 & \boldsymbol{\xi}^T\end{pmatrix}\begin{pmatrix}z_k-c_l-\boldsymbol{\beta}_l^T\mathbf{x} & \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x})^T\\ \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x}) & \mathbf{Z}_k-\mathbf{A}_l\end{pmatrix}\begin{pmatrix}1\\ \boldsymbol{\xi}\end{pmatrix}\\
& \geq 0 \ \ \forall\boldsymbol{\xi}\in\Omega_k,\\
& \forall l=1,...,L,\ \forall k=1,...,r,\\
& z_k+{\mathbf{z}_k}^T\tilde{\boldsymbol{\mu}}_k+\mathbf{Z}_k\cdot(\tilde{\boldsymbol{\Sigma}}_k+\tilde{\boldsymbol{\mu}}_k\tilde{\boldsymbol{\mu}}_k^T)-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\\
& \forall k=1,...,r,\\
& \gamma_k\geq(\boldsymbol{\beta}_l+\mathbf{D}_l\mathbf{m}_k)^T\mathbf{x}+\mathbf{m}_k^T\mathbf{A}_l\mathbf{m}_k+\mathbf{b}_l^T\mathbf{m}_k+c_l,\ \forall\, l=1,...,L,\\
& \forall k=r+1,...,K,\\
& \gamma_k-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\ \forall k=r+1,...,K,\\
& \lambda\geq 0.
\end{aligned}
\tag{3.5}
$$

Chapter 4

Algorithm

We can formulate RF as an SDP or SCP for different types of fragmentations as described in the previous section. But how do we obtain the parameters in the optimization problem and the regions in the fragmentation? In practice, the moments are unknown and must be estimated from training data. In classical MDRO, this is generally ignored as true moments are assumed to be known. In application, moments must be estimated, but estimation error is assumed to be negligible. We do not ignore this error. In fact, it is an important consideration for us in terms of building the ambiguity set. If we assume exact information about moments, we are always better off with a finer partition; but then we may as well assume we know the entire distribution. In truth, as we hone in our regions to get more detail, the estimates will be less accurate, a tradeoff we must balance. Our purpose is to estimate the structure of the distribution without using too fine a grain, so that we may dissect the crucial information, but not overfit to the training sample. We use this to guide the size of the regions we construct. The shape of each region will be dependent upon considerations of tractability. The size of the training sample and the modal structure of the distribution will help inform our fragmentation. In the following section, we discuss a procedure for obtaining a fragmentation over which to compute parameter estimates.

4.1 Implementation

We begin with a training sample. From the data, we must fragment and then estimate conditional moments. Given a fragmentation, we simply use the sample means $\hat{\boldsymbol{\mu}}_k = \frac{1}{N_k}\sum_{\boldsymbol{\xi}_i \in \Omega_k} \boldsymbol{\xi}_i$ and sample covariances $\hat{\boldsymbol{\Sigma}}_k = \frac{1}{N_k - 1}\sum_{\boldsymbol{\xi}_i \in \Omega_k}(\boldsymbol{\xi}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{\xi}_i - \hat{\boldsymbol{\mu}}_k)^T$, where $N_k > 1$ is the number of data points contained in region Ω_k, to get our estimates for the conditional moments. We use the sample values because they are consistent and unbiased estimators. Consistency guarantees us convergence of the optimal value and optimal solution as sample size $N \to \infty$, as detailed in Section 2.5. We could, in theory, choose any consistent estimators to ensure this result. We choose unbiased estimators so that we do not introduce a bias in the estimates that is carried through to the solution of the optimization problem. In some cases it may make sense to use more robust estimators, such as in very high dimensions or in cases with many outliers.

A key motivation of the fragmentation approach is that it captures modal information that is unavailable to MDRO. In order to utilize this additional information, we need to be able to identify modes of the distribution. Finding the modes of a distribution based on a sample of points is a challenging problem and is akin to clustering the data points based on certain criteria.

We focus on a couple of simple approaches to finding distributional modes based on clustering algorithms. In principle, other methods could be substituted and were considered, but each would have advantages and drawbacks. In Chapter 5, we discuss in detail a variety of techniques to consider and how they compare in terms of different performance criteria. In general, any method should be applied with care and consideration for the specific problem at hand. Our objectives include minimizing the estimation errors in the regional moments and reducing instability by encouraging higher conditioning numbers of regional ellipsoidal shape matrices. These must be balanced with our other objectives of accurately capturing the modality of the data and minimizing reconstruction error, which may be in conflict.

Proposition 4. *Given a fixed number of regions r, estimation errors in the conditional moments will be minimized by assigning $\frac{N}{r}$ points to each region (ignoring remainders).*

Proof. By the central limit theorem, we know that the errors in both the regional sample means $\hat{\boldsymbol{\mu}}_k$ and the regional sample covariances $\hat{\boldsymbol{\Sigma}}_k$ scale asymptotically as $\frac{1}{\sqrt{N_k}}$. To

minimize the errors, we solve for $\min_{\{N_k\}} \sum_k \frac{1}{\sqrt{N_k}}$ subject to $\sum_k N_k = N$. We consider the Lagrangian function and solve for the KKT conditions by setting the gradient of the Lagrangian equal to zero as follows:

$$\nabla(\textstyle\sum_k \frac{1}{\sqrt{N_k}} + \lambda(\sum_k N_k - N)) = \begin{pmatrix} \vdots \\ -\frac{1}{2N_k^{\frac{3}{2}}} + \lambda \\ \vdots \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix},$$

where λ is the Lagrange multiplier. Therefore we have that $-\frac{1}{2N_k^{\frac{3}{2}}} = -\frac{1}{2N_l^{\frac{3}{2}}}$ $\forall k, l$, and thus all of the N_k are equal. Since $\sum_k N_k = N$, this implies that $N_k = \frac{N}{r}$ for all k. This gives us a general guideline that spreading out the estimation errors will help minimize the total estimation errors. □

Additionally, $\hat{\boldsymbol{\Sigma}}_k = \frac{1}{N_k - 1} \sum_{i=1}^{N} I_{ik}(\boldsymbol{\xi}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{\xi}_i - \hat{\boldsymbol{\mu}}_k)^T$, where $\hat{\boldsymbol{\mu}}_k$ is the estimated mean of cluster k. Thus, we need $N_k = \sum_{i=1}^{N} I_{ik} \geq p + 1$ in order for $\hat{\boldsymbol{\Sigma}}_k$ to be nonsingular (since it is a sum of rank 1 matrices only $N_k - 1$ of which can be linearly independent). If the distribution is continuous, then there is zero probability that they will be linearly dependent (assuming the support is not a set of measure zero). For a discrete distribution, the more points in the sample, the greater the rank of the covariance matrix. Singular or close to singular covariance matrices cause instability in the method since the ellipsoidal regions will be very "flat", with small perturbations in points causing large changes in the regions. Singularity will also cause numerical instability in the algorithm. Therefore, in general, evening out the number of points in all the regions will result in the most accurate estimates and highest stability.

Although evening out the points will improve estimation and stability, we must balance this with trying to find the modes of the distribution. To this end, we develop an adaptation of k-means that includes a penalty term for every cluster that does not contain enough points. The threshold number of points τ_k for cluster k may be based on either of the previous two arguments we have discussed. Thus, it could be given as $\tau_k = p + 1$ to ensure that each regional covariance matrix is nonsingular (in continuous case) with probability 1, or it could be given as $\tau_k = \frac{N}{k} - \zeta$ where ζ is an allowed leeway from requiring all clusters to be of equal size. Additionally, this method involves a tuning parameter α which determines the size of the penalty for clusters that have

too few points. This constrained version of k-means may be solved using an adaptation of Lloyd's algorithm, where in each assignment phase a linear program must be solved. This is will be discussed in detail in Chapter 5.

Another effective method for clustering the data sample is a Borgelt and Kruse (2005) adaptation of the well-known expectation-maximization algorithm in application to Mixture of Gaussians clustering. Under this model, we assume that the data follows a mixture distribution where each component is Gaussian. Given the number of components r, we can use a maximum-likelihood approach to find parameter values. The likelihood function is not tractable so hidden variables are introduced to represent component responsibilities. The complete log-likelihood is maximized by employing the expectation-maximization algorithm which was first developed in Dempster et al. (1977). Once the algorithm has converged to an optimal set of parameter values, we may compute the densities of each data point under each component and assign points to groupings based on maximum density. Borgelt and Kruse (2005) introduce an additional step in each iteration, where the regional covariance estimates are updated by shifting all of the eigenvalues up by some specified amount and then renormalized to the same determinant. This is equivalent to adjusting the axes of the corresponding ellipsoids to be more hyper-spherical to reduce the occurrence of clusters with long, thin ellipsoids or with very few points. Thus, the adaptation will lead to higher stability in the ellipsoid shapes and the algoritm and spread out the points more evenly between modes. Each group is considered a mode of the distribution and corresponds to a region in the fragmentation. Modes will generally be Gaussian shaped, at least as much as the data will allow, as points will be assigned based on a fitted Gaussian component. This fits very naturally into an ellipsoidal fragmentation, as level curves of Gaussian densities are ellipsoids. By constructing minimum volume containing ellipsoids for each cluster, we get an ellipsoidal fragmentation that is mostly disjoint and covers all points in the sample. Additionally, we may further divide each ellipsoid into differences of ellipsoids according to the number and distribution of sample points inside the region.

Let us be clear that we are not assuming that the distribution of Ξ is a Mixture of Gaussians. We are still retaining the distributional ambiguity set defined by the conditional moments constraints. We are, however, deriving these constraints based on a partition constructed from the MOG clustering approach. We are only using it as a

tool to identify the modal structure.

We employ both of these methods for modal detection depending on the application and domain knowledge. Additionally, we performed some experiments with other clustering methods, including density-based spatial clustering of applications with noise (DBSCAN), which is effective for nonconvex clusters, introduced by Ester et al. (1996). All clustering methods used will be illustrated and compared in Chapter 5. After clustering, we construct minimum volume ellipsoids containing each subsample. Within each ellipsoid, we may break up into differences of ellipsoids depending on the subsample size.

Both clustering methods discussed so far require r as an input. Thus, we need to determine the number of regions r in the fragmentation. There are two factors affecting this decision: sample size and number of modes in the data. As N grows, so should r to retain the same optimal level of robustness. Fragmentation is designed to capture modal information, so the number of modes should play an important role as well. First we will discuss the relationship between r and N.

We can use a bound on the difference between optimal values to help determine how the number of regions should scale with the number of data points. Suppose that $\pi^\star$ is the true, unknown distribution for $\tilde{\boldsymbol{\xi}}$. Suppose the support Ω of $\boldsymbol{\xi}$ is compact and we partition into r equally sized regions of diameter d (no singletons). We have a data sample of size N. We assume the cost function $f(\mathbf{x}, \boldsymbol{\xi})$ is bounded by M and Lipschitz with constant C. We consider the difference between the worst case expected cost Υ_N and the expected cost $\Upsilon^\star$ over the true distribution $\pi^\star$. The difference between objective function values $|v_N(\mathbf{x}) - v^\star(\mathbf{x})|$ and optimal values $|\Upsilon_N - \Upsilon^\star|$ measure how closely we are able to approximate the stochastic program.

Theorem 5. *We have that,*

- *a)* $\forall \epsilon > 0$, $\lim_{N \gg r \to \infty} \mathcal{P}\left(|\Upsilon_N - \Upsilon^\star| \geq \epsilon\right) = 0$, *and*
- *b)* $\mathcal{P}\left(|\Upsilon_N - \Upsilon^\star| \geq M\gamma(r, N) + M\frac{\Phi_{1-\frac{\alpha}{2}}}{\sqrt{N}}\sqrt{r-1} + C'r^{-\frac{1}{p}}\right) \lesssim \alpha$ *for some classes of* ϕ*-divergence functions.*

Proof. The proof is given in Appendix B.3. □

The first of the three terms in the bound in part b) is what we refer to as the "probability error". This term bounds the difference in optimal values due to the ambiguity set for the probability vector. As $N \to \infty$, this term vanishes ($\rho \to 0$) as the need for ambiguity shrinks with increased accuracy. The second term in the bound is related to errors in the moment estimates and can be thought of as the "estimation error", which increases with the number of regions r, but decreases with sample size N. The last term is error caused by using the robust method: even if the moments are correct, the solution will still not be optimal for the true distribution. We call this the "conservative error", which decreases as the regions become smaller and the ambiguity set is reduced.

We can use this equation to gain insight into how r should scale with N, in order to minimize the difference between the estimated optimal value and the true expected cost, by taking r as the minimizer of the upper bound on the error in the second line. For instance, for variation distance, we have $\min_{r \in \mathcal{N}} \mathcal{O}(\frac{1}{N}) + \mathcal{O}(\frac{\sqrt{r}}{\sqrt{N}}) + \mathcal{O}(r^{-\frac{1}{p}})$. We set $\frac{d}{dr} = \mathcal{O}(\frac{1}{\sqrt{r}\sqrt{N}}) - \mathcal{O}(r^{-\frac{1}{p}-1}) = 0$. We arrive at $r \propto N^{\frac{p}{p+2}}$. Thus, for $p = 1$, $r \propto N^{\frac{1}{3}}$, for $p = 2$, $r \propto \sqrt{N}$, and as $p \to \infty$, $r \propto N$. As the dimension grows, r should scale more rapidly with N. This holds true for other ϕ-divergences as well, with the first term bounded above in the worst case by $\mathcal{O}(\frac{\sqrt{r}}{\sqrt{N}})$.

This helps inform how r should scale with N and we performed some numerical experiments to further explore this relationship. We looked at out-of-sample performance for the newsvendor problem (with $p = 2$) in terms of the training sample size N for MDRO. The newsvendor problem and our setup is described in greater detail in Section 6.1. There is no probability ambiguity and the conservative error is constant. We focus on how the estimation error depends on N. Our goal is to understand how more accurate moments estimates translate to a more accurate solution to the optimization problem. Recall that the estimation errors scale as $\frac{1}{\sqrt{N}}$. In Figure 4.1, we construct a log-log plot of the relative expected cost vs. sample size N. The relative, or additional cost, refers to the difference between the average expected cost at sample size N and the expected cost for infinite sample size (moments are exact). In this case, the slope of the line indicates that the (Expected Cost - Infinite Sample Cost) scales as $\frac{1}{N^{0.7235}}$. This tells us that there are decreasing returns to scale for increased sample size N. At some point, more sample points provide little benefit in terms of moments accuracy and we

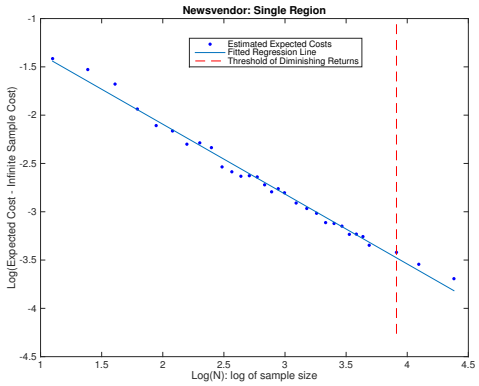

Figure 4.1: Log-Log Relative Cost vs. N

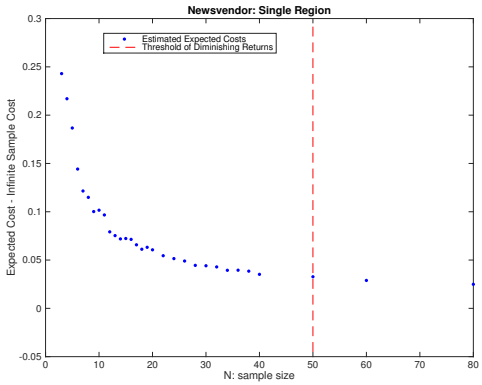

Figure 4.2: Diminishing Returns for MDRO

will be better off breaking up into smaller regions to reduce the conservative error. We can see how the return diminishes on the right in Figure 4.2. This is just one of many experiments demonstrating this concept. As a heuristic we estimate that a subsample size of $N = 50$ is sufficient for each region.

Although we only show an example with Gaussian data here, this result is more general. We tested this with quadrimodal data and using robust fragmentation with four regions to see if the same trend occurs with multiple regions. In Figure 4.3, we can see that the performance levels off for N greater than 200, which equates to just over 50 points per region. These results indicate that additional points beyond 50 per region do not improve moments accuracy enough to generate additional improvements in performance. In Figure 4.4, we can see the optimal r value for different values of N, where in this case the data is uniform and the fragmentation is a polytopal partition. For $N = 40$, $r = 4$ is best; $r = 16$ for $N = 400$; and $r = 64$ for $N = 4000$. The ratio $\frac{\sqrt{N}}{r}$ stays fairly constant (considering the discrete nature of the values chosen). The optimal number of points per region is less than 50 because the data is uniform rather than a mixture of gaussians, and has no modal tendencies. We conclude, from these demonstrations and many others not presented here, that an optimal r will depend on the structure of the data and the dimension, but that a factor of $\frac{N^{\frac{p}{p+2}}}{r} = 20$ to 50 is appropriate for many applications. If the distribution is multimodal, the number and location of regions should reflect this (for multimodal distributions, even small N will justify enough regions to cover the modes).

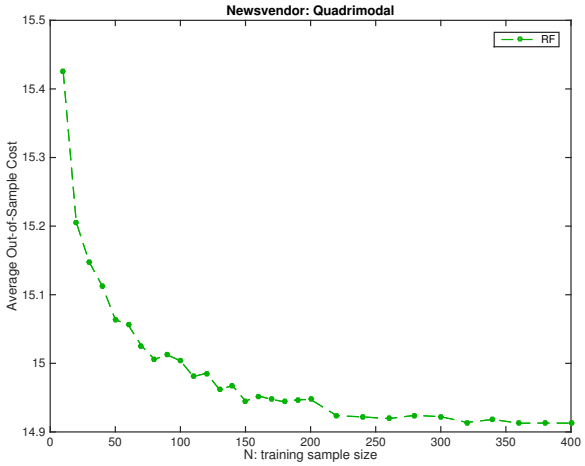

Figure 4.3: Diminishing Returns for RF with 4 regions

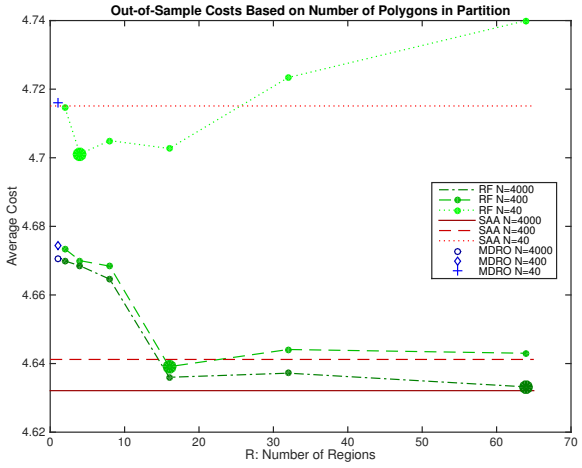

Figure 4.4: Optimal Number of Polytopes

As described in the previous few paragraphs, heuristics can give a general guideline for how to select r based on sample size N. However, these should be combined with knowledge of the number of modes in the distribution, if available. If there are clearly separated groups of points (known either through a priori knowledge or visualization in 2 or 3 dimensions), then r should be chosen as the number of groups, unless clusters have fewer than 20 points and are loosely formed, in which case they should be ignored. If modal information is unknown, there are estimation methods. There are many such preprocessing techniques designed for k-means or clustering in general, and we introduce one simple one here. Keep in mind that finding the exact number of modes (if even possible) is not our goal - only to have a general idea of an appropriate number for our purposes of optimization, which may not even be equal to the true number.

To this end, we employ what is called the "gap statistic" as introduced in Tibshirani et al. (2001). One measure of cluster performance is the within-cluster variance,

$$WCSS = \sum_{i=1}^{N} \sum_{k=1}^{r} I_{ik} \|\mathbf{x}_i - \mathbf{m}_k\|^2, \tag{4.1}$$

where $\mathbf{x}_i$ is the ith data point and

$$\begin{aligned} I_{ik} &= 1 \quad \text{if } \|\mathbf{x}_i - \mathbf{m}_k\| = \min_t \|\mathbf{x}_i - \mathbf{m}_t\|, \\ &= 0 \quad \text{else.} \end{aligned} \tag{4.2}$$

In k-means, the objective is to minimize $WCSS$ over all cluster assignments to r regions. As r increases, $WCSS(r)$ will decrease monotonically, so comparing values at different r is uninformative. Instead, we compare $WCSS(r)$ at each value of r relative to the $WCSS$ for the same r and N but under a uniform distribution. That is, we draw a sample of size N from a uniform distribution over the same support as our training sample and compute $WCSS(r)$ for different values of r for both distributions. For the uniform sample, the decrease will be due entirely to an effect of having more cluster centers and therefore smaller average distances. For the problem data, this effect will be mixed with the true effect of capturing the modality of the distribution. Thus, the relative difference between the two will capture the effect from the modality. Suppose there are a true number of modes m. As long as $r < m$, decreases in $WCSS(r)$ should

be greater for the training sample than for the uniform sample. As $r > m$, decreases in $WCSS$ will be less pronounced than for smaller values of r and the gap between the two $WCSS$ values will decrease. We choose a value for r that maximizes the gap to best approximate the true number of modes.

We combine estimates of the number of modes with requirements for the number of points in each mode. Any clusters resulting in fewer than some threshold number of points (e.g. 20), are dissolved into single points and put into the summation term in (2.3). If too many of the groups are too large, we will choose a larger value for r and if too many clusters are being dissolved, we will decrease r. In the next chapter, we will discuss the clustering procedures in further detail.

Chapter 5

Clustering

Clustering is an unsupervised learning task to find structure in unlabeled data. It can be thought of as grouping objects, or data points, in some way such that the objects in each group are more similar to each other than to objects in other groups. Such groups are referred to as "clusters". Thus, the objective of clustering is to find an inherent grouping structure in the data. It has applications in almost every field you can think of, such as grouping customers, classifying plants, and discerning individuals who do not belong (outliers). We require clustering methods for grouping historical data as a preprocessing technique in robust fragmentation. We are interested in how various clustering methods perform in terms of the goals of the clustering and the overall performance of the optimization technique. To this end, we analyze a variety of different clustering methods as a part of robust fragmentation. We consider several well known methods including K-means, Mixture of Gaussians, and Density Based Spatial Clustering of Applications with Noise (DBSCAN). We also introduce some adaptations to some of these methods in an effort to achieve desired properties. We then compare performance of the methods in terms of various designed metrics, some of which are very general and others of which are tailored to our ultimate purpose. Each method makes different assumptions about what it means for objects (data points) to be similar. In our analysis, we attempt to understand the differences between these methods and how their ability to group data is affected by the structure of the data. We use several types of data generating models and additionally consider real financial data in order to observe how each method performs when assumptions are and are not met. We

make some observations about how the metrics change for each clustering technique as a function of the data type, dimension, and number of clusters. We restrict ourselves to a few simple techniques and extensions, so as not to overly complicate the procedure. We make some conclusions about preference of methods, but strongly emphasize that these insights are circumstantial, depending on the data.

5.1 Clustering Methods

We test different techniques for clustering data points in Euclidean space. We exploit a few well known methods, which we briefly explain in this section. We also expand upon some of these methods to get desirable qualities. We will focus on three main methods and adaptations, as described in the following three subsections.

5.1.1 K-means

The first, and most simple method, is known as the k-means algorithm which dates back to 1957 but was published by Lloyd (1982) 25 years later (although essentially the same idea appeared in other papers). The basic idea behind this algorithm is that we are going to compress our data into a fixed number of vectors, called "code vectors". K-means aims to minimize the differences between the original data and the set of code vectors. Every data point will be represented with the most similar entry of the K reference vectors, where similarity is measured by Euclidean distance. In order to calculate the reference vectors $\mathbf{m}_k$ for $k = 1, ..., K$, we assume that we want the vectors to represent the original data points as closely as possible. That is, we want to minimize the sum of the differences between the original data set and the compressed data set where each vector has been replaced by its reference vector. Thus we wish to minimize the following quantity:

$$\sum_{i=1}^{N} \sum_{k=1}^{K} I_{ik} ||\mathbf{x}_i - \mathbf{m}_k||^2, \tag{5.1}$$

where $\mathbf{x}_i$ is the ith data point and

$$\begin{aligned} I_{ik} &= 1 \quad \text{if } ||\mathbf{x}_i - \mathbf{m}_k|| = \min_t ||\mathbf{x}_i - \mathbf{m}_t||, \\ &= 0 \quad \text{else.} \end{aligned} \tag{5.2}$$

Unfortunately, since I_{ik} depend on the $\mathbf{m}_k$, this problem cannot be solved analytically. Instead, it is solved iteratively. In alternate steps, either the I_{ik} and the $\mathbf{m}_k$ are assumed to be known, and we solve for the other. If the I_{ik} are assumed to be known, then we can take the derivative of 5.1 with respect to $\mathbf{m}_k$ and set it equal to zero. We then solve to get $\mathbf{m}_k = \frac{\sum_{i=1}^{N} I_{ik}\mathbf{x}_i}{\sum_{i=1}^{N} I_{ik}}$. Thus, the reference vectors will be chosen as the means of all of the points that they represent. Once we have the means $\mathbf{m}_k$, for each i we may simply determine which mean $\mathbf{m}_t$ is closest and then set $I_{it} = 1$ and the rest equal to zero. We continue to repeat this process until the convergence criteria is met, which is stabilization of the means. In order to start this iterative procedure, we need some initial guesses for the means. This can be done in several ways. A couple simple ways are to choose K random data points as the starting points or find the center of the data and perturb slightly to construct K points close to the center. Since in each step the value of the objective function is reduced and there are only a finite number of assignment arrangements, convergence is guaranteed. However, since it is nonconvex, this convergence may be to a local minima. To combat converging to an undesirable minimum, k-means is repeated several times and the best solution taken. (Alpaydin (2014))

As we discussed in Chapter 4, in addition to accurately capturing the modality of the data and minimizing reconstruction error, we wish to minimize the estimation errors in the regional moments and reduce instability. Both estimation error and stability will be improved by evening out the number of points in each region. Traditional k-means provides no guarantee that the clusters will not be of very different sizes and possibly too small. We wish to control the size of the clusters without losing the true modality structure of the data. To this end, we introduce a new variation of k-means that enables more control over cluster size.

Constrained K-means

We adapt k-means to include a penalty for clusters that do not contain enough points. For each cluster k, the threshold number of points τ_k may be based on either of the previous two arguments we have discussed. Thus, it could be given as $\tau_k = p + u$ for every k, where $u \geq 1$, to ensure that each regional covariance matrix is nonsingular (in continuous case) with probability 1, or it could be given as $\tau_k = \frac{N}{k} - \zeta$ for every

k, where ζ is an allowed leeway from requiring all clusters to be of equal size. This method involves a tuning parameter α which determines the penalty size for clusters that have too few points. The idea behind the design is similar to a constrained version of k-means proposed by Bradley and Bradley et al. (2000), where they had hard lower bounds on the size of each cluster, whereas we have only penalties.

The layout is the same as for k-means, except with the additional penalty terms. The objective criteria we consider is as follows:

$$\begin{aligned} \min_{I,\mathbf{m}} \ & \sum_{i=1}^{N}\sum_{k=1}^{K} I_{ik}||\mathbf{x}_i - \mathbf{m}_k||^2 + \alpha \sum_{k=1}^{K} \left(\tau_k - \sum_{i=1}^{N} I_{ik}\right)_+, \\ \text{subject to } & I_{ik} = 0, 1 \quad \forall i, k, \\ & \sum_{k=1}^{K} I_{ik} = 1, \ \forall i = 1, ..., N. \end{aligned} \tag{5.3}$$

This can be reformulated as follows:

$$\begin{aligned} \min_{I,\mathbf{m}} \ & \sum_{i=1}^{N}\sum_{k=1}^{K} I_{ik}||\mathbf{x}_i - \mathbf{m}_k||^2 + \alpha \sum_{k=1}^{K} \beta_k, \\ \text{subject to } & I_{ik} = 0, 1 \ \ \forall i, k, \\ & \sum_{k=1}^{K} I_{ik} = 1, \ \ \forall i = 1, ..., N, \\ & \beta_k \geq \tau - \sum_{i=1}^{N} I_{ik}, \ \forall k = 1, ..., K, \\ & \beta_k \geq 0. \end{aligned} \tag{5.4}$$

We consider an adaptation of Lloyd's Algorithm (1982), the iterative algorithm used to solve k-means as described in the section on k-means. Similar to k-means, we use alternating steps to solve for the I_{ik} and the $\mathbf{m}_k$. For known assignments I_{ik}, solving for the $\mathbf{m}_k$ to minimize the objective function is exactly the same as before. The minimizers will occur at the means of each cluster group. Thus, this step does not change from k-means.

In the assignment phase, the problem is now more challenging. We must now jointly

solve for the β_k with the assignments I_{ik}. Problem (5.4) is now an integer program that does not have an obvious solution as before. We cannot simply assign each point to the nearest cluster mean, since that will induce a penalty of α if another cluster does not have at least τ points. In fact, we can think of the penalized problem as follows. This will be an intuitive explanation of the concept behind the clustering, although not how it is implemented. We refer to clusters with number of points $N_k \geq \tau_k + 1$ as abundant clusters, those with $N_k = \tau_k$ as sufficient, and those with $N_k < \tau_k$ as deficient clusters. For fixed $\mathbf{m}_k$, without the penalty term, we would assign each point to the cluster with the nearest center $\mathbf{m}_k$. If this results in a solution such that all clusters are sufficient, then this will be the solution. If there exists a deficient cluster k under this arrangement, then we consider adding the nearest point $\mathbf{x}_i$ to that cluster that does not currently belong to it and belongs to an abundant cluster $k^{'}$. Switching the point from one cluster to another would reduce the objective function by an amount of the reduced penalty, α, but also increase the objective by the difference in distances to cluster means, $||\mathbf{x}_i - \mathbf{m}_k||^2 - ||\mathbf{x}_i - \mathbf{m}_{k'}||^2$. Therefore, we switch clusters if $\alpha \geq ||\mathbf{x}_i - \mathbf{m}_k||^2 - ||\mathbf{x}_i - \mathbf{m}_{k'}||^2$. We continue in this way, adding points to cluster k until it is sufficient or $\alpha < ||\mathbf{x}_i - \mathbf{m}_k||^2 - ||\mathbf{x}_i - \mathbf{m}_{k'}||^2$ where $\mathbf{x}_i$ is the nearest point from the nearest abundant cluster k'. Note that we would not swap with a sufficient cluster, as the penalty terms would cancel out and there would be no gain. We continue this swapping procedure for all deficient clusters until each has been made sufficient or has no additional points that would decrease the objective value.

This description gives an intuition into the tradeoff between cluster size equality and minimizing within-cluster sum of squares, but the algorithm does not rely on any swapping procedure and is quite efficient. Consider problem (5.4). We wish to show that the cluster assignment subproblem in each iteration is a Minimum Cost Flow (MCF) linear network optimization problem.

An MCF problem is associated with an underlying network. The network consists of a set of nodes N and edges E between some nodes. Each node has an associated value b_i, either supply ($b_i > 0$) or demand ($b_i < 0$). The sum of all supplies must equal the sum of all demands, that is, $\sum_i b_i = 0$. The set of edges between nodes allow flow along them for a per unit cost c_{ij} and the amount of flow is given by variable F_{ij}. The objective is to minimize total cost while meeting all node demands. The general MCF

can be formulated as follows:

$$\begin{aligned} \min \quad & \sum_{(i,j)\in E} c_{ij}F_{ij}, \\ \text{subject to} \quad & \sum_j F_{ij} - \sum_j F_{ji} = b_i, \ \ \forall i \in N, \\ & 0 \le F_{ij} \le u_{ij}, \ \ \forall (i,j) \in E, \end{aligned} \tag{5.5}$$

where u_{ij} is an upper bound on the flow over each edge. The first set of constraints requires that the amount of flow into and out of each node equal the supply/demand.

Theorem 6. *The assignment phase of problem* (5.3) *is equivalent to a minimum cost flow (MCF) problem.*

Proof. We let each data point $\mathbf{x}_i$ be represented by a node with supply $b_i = 1$ and each cluster k be a node with demand $-\tau_k$. From this point forward, we shall assume that all τ_k are the same and equal to τ. This is not necessary in general, but since our goals of covariance stability and minimal estimation error do not differentiate between clusters, there is no need for us to consider separate τ_k. We have a set of arcs (i,k) from each data point to each cluster with costs $c_{ik} = ||\mathbf{x}_i - \mathbf{m}_k||^2$. We then introduce an end node, call it node Z, with demand $-N$, with edges running from all cluster nodes to it. Additionally, we have an extra supply node, call it node S, with supply $k\tau$, with edges with cost α running from it to each cluster node and an edge with no cost running to the final demand node. The upper bound $u_{ik} = 1$ for all i,k, and $u_{kZ} = N$ for all k, while $u_{Sk} = \tau$ for each k and $u_{SZ} = k\tau$. This situation is depicted in Figure 5.1.

Now we aim to show that the solution to this MCF problem is in fact the same as the solution to the assignment problem. Consider the objective function:

$$\begin{aligned} \sum_{(i,j)\in E} c_{ij}F_{ij} &= \sum_{(i,k)\in E} c_{ik}F_{ik} + \sum_{(k,Z)\in E} c_{kZ}F_{kZ} + \sum_{(S,k)\in E} c_{Sk}F_{Sk} + c_{SZ}F_{SZ} \\ &= \sum_{(i,k)} F_{ik}||\mathbf{x}_i - \mathbf{m}_k||^2 + \sum_k \alpha F_{Sk}. \end{aligned} \tag{5.6}$$

If we let $F_{ik} = I_{ik}$ and $F_{Sk} = \beta_k$, then this is an equivalent objective function to that in (5.4) (we will see that the constraints will also be identical on these variables).

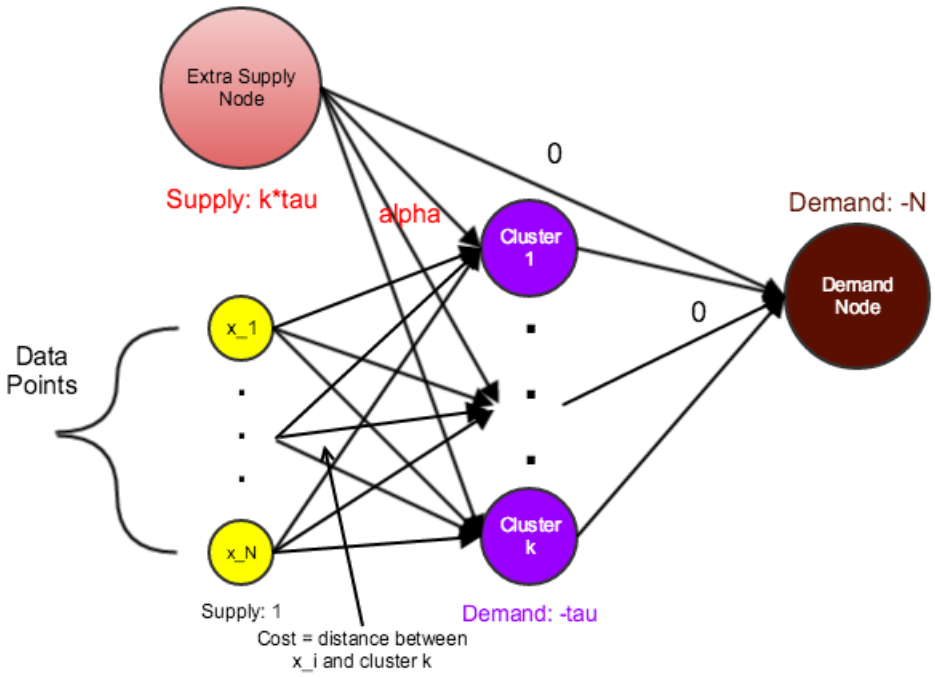

Figure 5.1: MCF Network

Consider the constraints $0 \leq F_{ij} \leq u_{ik} = 1$ for any i, k. For an MCF with integer constraints, the relaxation is equivalent to the integer program, which would have constraints $I_{ik} = 0, 1$, identical to those in (5.4). For each supply node corresponding to $\mathbf{x}_i$, we have the flow constraint $1 = \sum_{k=1}^{K} F_{ik}$, which is equivalent to the constraints $\sum_{k=1}^{K} I_{ik} = 1, \ \ \forall i = 1, ..., N$. For supply node S, we have that $k\tau = \sum_{k=1}^{K} F_{Sk} + F_{SZ}$, so that $F_{SZ} = k\tau - \sum_{k=1}^{K} F_{Sk}$. Once again, we replace each F_{Sk} with β_k and arrive at the constraint $F_{SZ} = k\tau - \sum_{k=1}^{K} \beta_k$. For the final demand node, we have the flow constraint $-N = -\sum_{k=1}^{K} F_{kZ} - F_{SZ}$. We can combine these two constraints, eliminating the variable F_{SZ}, to get that $N = \sum_{k=1}^{K} F_{kZ} + k\tau - \sum_{k=1}^{K} \beta_k$. Finally, for each node corresponding to a cluster k, we have flow constraint $-\tau = -\sum_{i=1}^{N} F_{ik} - \beta_k + F_{kZ}$.

This last constraint is equivalent to $\beta_k = -\sum_{i=1}^{N} F_{ik} + \tau + F_{kZ}$. Since $F_{kZ} \geq 0$, this implies that $\beta_k \geq \tau - \sum_{i=1}^{N} F_{ik}$, where we can drop the variable F_{kZ}, since it simply represents the slack in this inequality. This slack can be thought of as the extra points in cluster k, as if there are extra points, then the penalty size $\beta_k = 0$ and $F_{kZ} = \sum_{i=1}^{N} F_{ik} - \tau$. The constraint $N = \sum_{k=1}^{K} F_{kZ} + k\tau - \sum_{k=1}^{K} \beta_k$ simply tells us that if we add together the total required number of points in clusters, $k\tau$, with the extra points in clusters, and then we subtract the total number of penalties (points under the required number), then we get the total number of points, N. The bound $0 \leq F_{Sk} \leq \tau$ $\forall k$ requires that the number of points β_k lended to any cluster be less than τ, which will always be true. The bound $0 \leq F_{kZ} \leq N - \tau$ constrains the number of extra points in an abundant cluster be no more than $N - \tau$. Meanwhile, the constraints $\beta_k \geq \tau - \sum_{i=1}^{N} I_{ik}$ are the same as in (5.4). Therefore the two formulations are the same. Note that we assume that τ is an integer and also that $k\tau \leq N$ so that the problem is feasible. $\square$

Now we have shown that the assignment problem is in fact an MCF. We did this because then we can relax the constraints on the assignments to $0 \leq I_{ik} \leq 1$, and solve a linear program instead of the more difficult integer program, without changing the solution. Bertsekas showed that any MCF with all integer constraints has an integer solution (1991). Therefore, it is guaranteed that even if solving the relaxed linear program, the cluster assignment step will place each point in exactly one cluster. For (5.3), all of the demands and supplies are integer - τ, N, 1, and $k\tau$, as well as the upper bounds, $u_{ij} = 1, \tau, N - \tau, k\tau$. Therefore, the solution will always be integer valued by Proposition 2.3 of Bertsekas (1991). Thus, this linear program MCF will have the same

solution as the integer program MC, and therefore the same solution as (5.4).

Proposition 5. *The proposed adaptation of Lloyd's algorithm to solve* (5.3)*, alternating between computing centroids and assigning points to clusters by solving the MCF assignment problem in each assignment step, will converge in a finite number of steps to a local minimum.*

Proof. Note that for each step of computing the means $\mathbf{m}_k$, the objective value cannot increase, just as in regular k-means. This is because we are minimizing a convex function (for fixed assignments) and so will always get the optimal centers for each cluster. In the assignment phase, we also cannot increase the value of the objective function. Given the new centers, the solution to the integer program, which we can solve exactly, will minimize the objective. Therefore, the objective value cannot increase in each step. Since there are only a finite number of possible assignments, we must eventually reach a point where the assignments do not change in an iteration. If this is the case, then the centers will not change in the following iteration, and we have reached a stationary point. This must be a local minimum, since if it were not, then one of the two steps would not be stationary. Since it is a network problem, we can run codes specifically tailored to network optimization which run even faster than ordinary LP codes. □

Thus, our adaptation of k-means can be solved through iterative convex minimizations and solving minimum cost flow problems. It will also help in minimizing estimation errors and decreasing instability due to covariance singularities. The only issue is how to pick α, the penalty parameter. Thinking back to the swapping explanation for how points are assigned to clusters based on penalty size α gives a guideline for the size of the parameter. α should be larger than the minimum pairwise distance between points in the dataset because otherwise there will no incentive to swap any points. However, α should be smaller than the maximum pairwise distance between points or all clusters will be forced to be of size greater than or equal to τ, becoming a hard constraint. Therefore, α should be somewhere between these two extremes. Consider the average pairwise distance between points in the dataset. If we let α be some multiple of this average distance, then we would add a point to a deficient cluster with a further center point if the center is not further by more than a multiple of the average pairwise distance. This is a reasonable approach and we find that a factor of $\frac{1}{4}$ to $\frac{1}{2}$ works well to

discourage small clusters without forcing equal sized clusters.

5.1.2 Mixture of Gaussians

The next algorithm that we discuss, Mixture of Gaussians (MOG), can be thought of in some sense as a generalization of k-means, where the same iterative procedure is used to find cluster centers and assign points to clusters, but the assignments are "soft" in each step and are based on Gaussian density functions. Also, the philosophy behind the approach is very different as it is probabilistic. We assume that the data follows a distribution described by a mixture of Gaussians and estimate the parameters based on a maximum likelihood approach that requires the same kind of alternating step algorithm used in k-means. This algorithm is called expectation-maximization (EM) and was originally developed by Dempster et al. (1977). The assumption is that the data follows a mixture density written as:

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}|G_k)P(G_k), \tag{5.7}$$

where $p(\mathbf{x}|G_k)$ are the densities for each component distribution and $P(G_k) = \pi_k$ are the component probabilities. Each component is a Gaussian with mean $\mathbf{m}_k$ and covariance matrix $\mathbf{S}_k$. We want to find parameter values Ω to maximize the log-likelihood given by $L(\Omega|\mathbf{X}) = \sum_i \log \sum_{k=1}^{K} \pi_k p(\mathbf{x}_i|G_k)$. This cannot be solved analytically to maximize the likelihood of the observed values, $\mathbf{X}$. But we can introduce extra hidden, or latent variables $\mathbf{Z}$, and express the complete likelihood as a function of both $\mathbf{X}$ and $\mathbf{Z}$. The hidden variables $\mathbf{z}_i$ correspond to the responsibility of each component G_j for data point $\mathbf{x}_i$. That is, whether point i came from each component. We estimate the $\mathbf{z}_i$ by taking the expected values of the latent variables $\mathbf{Z}$ based on the current component parameters and then use these expected values to find a new set of component parameters that maximize the complete log-likelihood. In this way, the algorithm is very similar to k-means in that it alternates between steps of finding component means (and in this case, variances),and assigning points to clusters (finding the expected values). (Bishop (2006))

The two steps of this algorithm are called the maximization and expectation steps.

In the E-step (expectation), we compute the expected values for z_k^i, denoted by h_k^i, the responsibility of component k for point i. We plug these in to compute the expected complete log-likelihood, given the current parameter estimates:

$$\sum_i \sum_k E[z_k^i|\mathbf{X}](\log \pi_k + \log[p_k(\mathbf{x}_i)]) = \sum_i \sum_k h_k^i(\log \pi_k + \log[p_k(\mathbf{x}_i)]). \tag{5.8}$$

In the M-step (maximization), we maximize the expected complete log-likelihood to solve for the next set of parameter values. The maximum likelihood estimates are:

$$\pi_k = \frac{1}{N}\sum_i h_k^i, \tag{5.9}$$

$$\mathbf{m}_k = \frac{\sum_i h_k^i \mathbf{x}_i}{\sum_i h_k^i}, \tag{5.10}$$

$$\mathbf{S}_k = \frac{\sum_i h_k^i (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T}{\sum_i h_k^i}. \tag{5.11}$$

where the estimates for the means in the first equation are used to estimate the covariances in the second.

Thus, the full algorithm is as follows:
(1) Initialize means $\mathbf{m}_k$ using a couple iterations of k-means.
(2) Initialize probabilities π_k and covariances $\mathbf{S}_k$ by assigning points to closest mean.
(3) Continue with expectation step, and alternate from them on.
(4) Repeat until convergence (expected complete log-likelihood stabilizes). (Alpaydin (2014))

Constraining Cluster Shapes

Similar to how we adjusted k-means to prevent singularity of regional covariance matrices, we also consider an adaptation to the EM algorithm for MOG that encourages stability as well. We implement an extension of EM for MOG developed by Borgelt and Kruse (2005). They constrain the shape of the clusters by regularizing the covariance matrices in each maximization step. This consists of shifting the eigenvalues away from zero to make the ellipsoid more spherical, and preventing extremely skinny ellipsoids.

This is done without changing the orientation (eigenvectors) or the size (determinant) of the ellipsoid.

Let $\mathbf{1}$ be the identity matrix. Let $\sigma_k^2 = \sqrt[p]{|\mathbf{\Sigma}_k|}$ be the equivalent isotropic variance (leads to same hypervolume) for cluster k. Let $\mathbf{S}_k = \sigma_k^{-2}\mathbf{\Sigma}_k$ be a rescaling of the kth covariance matrix to have determinant 1. Our regularization parameter will be h^2. We only want to alter the shapes of covariances if they are singular or close to being singular. To this end, we consider a limit on the ratio of the longest major axis to the shortest major axis of each hyperellipsoid and only implement the adaptation if this limit is surpassed. The limit is determined by a parameter r^2. If $\frac{\max_j \lambda_j}{\min_j \lambda_j} \geq r^2$, then we implement the adaptation, where λ_j are the eigenvalues of $\mathbf{\Sigma}_k$. If this is the case, then we let $h^2 = \frac{\max_j \lambda_j - r^2 \min_j \lambda_j}{\sigma_k^2(r^2-1)}$, and otherwise let $h^2 = 0$. In each maximization step, we alter the covariance calculation by $\tilde{\mathbf{\Sigma}}_k = \sigma_k^2 \frac{\mathbf{S}_k + h^2\mathbf{1}}{\sqrt[p]{|\mathbf{S}_k + h^2\mathbf{1}|}}$. Thus, if $h^2 = 0$, then there is no change, otherwise a tendency towards hyperspherical clusters is introduced.

5.1.3 DBSCAN

The final clustering approach that we will describe is a density-based method. It is called Density Based Spatial Clustering of Applications with Noise (DBSCAN). It was first introduced by Ester et al. (1996) and is now one of the most highly cited clustering algorithms in the field of computer science. DBSCAN builds notions of connectivity between points to construct clusters. It is a density-based clustering algorithm that generates clusters according to whether an estimated density of the generating distribution is above a certain threshold at any given point. Thus, clusters exist in areas where points are dense. For a collection of points to be considered a cluster, the density of points in a neighborhood must pass a threshold.

The density is computed according to two parameters: ϵ, which defines a neighborhood size, and *minpts*, which defines the number of required neighbors. There are effective heuristics for determining these two parameters or in some applications it may be possible to use domain knowledge to inform their values. In our case, we simply set them based on our knowledge of the sample size, expected or desired number of clusters, and size of the support.

The algorithm is very scalable, iterating through points with an average run time of $\mathcal{O}(N \log N)$. Points are assigned uniquely up to "border points" between clusters.

Some points will not belong to any cluster and will be labeled as outliers/noise. For our purposes, these outliers would be incorporated as degenerate regions and put into the summation term over S in (2.3). The number of clusters r is not required a priori, which is a distinct advantage over the previous methods we have described. Another desirable characteristic is that there are no distributional assumptions, such as in MOG. The algorithm can handle arbitrarily shaped clusters, but this is not necessarily an advantage since the regions in the partition will necessarily be ellipsoidal, so we may not want to construct convex clusters, as this will cause more overlap in the fragmentation. DBSCAN also performs poorly with connected clusters or clusters of varying densities. (Ester et al. (1996))

5.2 Metrics for Evaluation

We use metrics to measure performance according to certain criteria. Each metric may be biased towards certain methods and will only measure one aspect of an effective grouping mechanism for our purposes. One criteria we use to evaluate the clustering algorithm is the within-cluster sum of squares. We compute the within-cluster variance, the minimization criteria for k-means, for each method as follows:

$$WCSS = \sum_{i=1}^{N} \sum_{k=1}^{r} I_{ik} \|\mathbf{x}_i - \mathbf{m}_k\|^2, \tag{5.12}$$

where $\mathbf{x}_i$ is the ith data point and

$$\begin{aligned} I_{ik} &= 1 \quad \text{if } \|\mathbf{x}_i - \mathbf{m}_k\| = \min_t \|\mathbf{x}_i - \mathbf{m}_t\|, \\ &= 0 \quad \text{else.} \end{aligned} \tag{5.13}$$

This is the sum over all points of the distance to the center of the cluster with which it has been identified. This gives us a measure of total variation inside of each cluster, summed over all clusters. It measures the loss of information incurred when reducing the data. The total sum of squares (TSS) of the points from the overall mean is given as the sum of the $WCSS$ and the weighted sum of squared distances to cluster centers (CSS), $TSS = WCSS + CSS$. The lower the $WCSS$, the higher the proportion of

TSS that is comprised of CSS. Since the data is reduced to the cluster centers, the higher the CSS, the higher proportion of the variation that is retained after clustering. Therefore, the higher the CSS, and thus lower the $WCSS$, the less information is lost in the reduction. This criteria favors k means, as it is exactly the criterion the k-means algorithm seeks to minimize. Since DBSCAN does not have a fixed number of clusters, we adjust this metric by multiplying the number of clusters k. With more clusters, the sum of squares will be artificially smaller (extreme case is where each point is a cluster and sum of squares is zero), so we multiply by k to penalize for the number of clusters (note that this primarily penalizes the density-based approach, as it often has a larger number of clusters).

We adapt a metric called the Davies-Bouldin Index (DBI), introduced by Davies and Bouldin (1979), to try to measure the average overlap between regions. The standard DBI is a completely internal metric (does not rely on labels). It is given by:

$$DBI = \frac{1}{K}\sum_{k=1}^{K}\max_{j\neq k}\left(\frac{\sigma_k + \sigma_j}{d(c_k, c_j)}\right), \tag{5.14}$$

where K is the number of clusters, σ_j is the radius of cluster j, and $d(c_k, c_j)$ is the distance between clusters k and j. Normally the radius is the mean distance from the centroid of cluster j and the distance between two clusters is the distance between the two centroids. However, since we are concerned with ellipsoidal regions and not spherical ones, we measure the radius as the average Mahalanobis distance from the center of the minimum volume containing ellipsoid for each cluster k, where the shape matrix is already determined. The distance between two clusters is given by $d(c_k, c_j) = \frac{d^\star(c_j,c_k)+d^\star(c_k,c_j)}{2}$ where the two distances $d^\star$ are computed using the shape matrix of clusters k and j, respectively. Thus, each k term in the sum can be thought of as the maximum overlap from another cluster with k. The "overlap" is the sum of the Mahalanobis radii of the two clusters divided by the distance between them. If this is 1, for example, we would expect a reasonable amount of overlap between the two clusters. As this value goes higher, the two clusters overlap more. For each cluster, we consider the maximum overlap and then average over all clusters. This gives us an average maximal overlap. The lower this value, the better we consider our algorithm in terms of this criteria. Our objective in this case would be to have low intra-cluster

distances and high inter-cluster distances (well separated clusters) and thus a low DBI. (Davies and Bouldin (1979))

Additionally, we record a few other measurements. The minimum eigenvalue of each covariance matrix is calculated to measure instability. The number of clusters with minimum eigenvalues below a threshold $\tau = .01$ is kept track of. We compare the computational time required for each method, added together for both the clustering portion, and for the cvx implementation, both in MATLAB R2014b. This measures how scalable the algorithm is and also how the cluster shapes it constructs affect the scalability of the interior point methods used to solve the resulting RF formulation. This will be affected by the number and shape of the clusters (more singularity causes more instability and slower convergence in cvx). We mention some other advantages/disadvantages of each method, such as the number of parameters that are required as inputs and whether or not it is able to deal with outliers.

The most important metric is the performance of the resulting RF implementation. This measures how well the clustering algorithm performs as a preprocessing technique in RF, and is our ultimate criteria. We looked at out-of-sample performance for the newsvendor problem (with $p = 2$) in terms of the clustering technique chosen. The newsvendor problem and the specific setup we consider will be described in more detail in Section 6.1. The important thing to keep in mind is that we are measuring out-of-sample costs, so that lower values are better. We do not consider probability ambiguity.

5.3 Experiments

To get an idea of the relative advantages and disadvantages of each clustering method, as well as the types of data for which it is applicable, we generate the data in four different ways and use real data. We attempt to meet the assumptions of each clustering method in one of the generating models we use. This is so we can see if the performance of each clustering method is best (relative to the others) when assumptions are met and also how each one generalizes when assumptions are not met.

5.3.1 Data-Generating Models

We will evaluate the four clustering methods on four different data models, each chosen to fit the assumptions of one of the clustering methods. In all generating methods, the user can control N, the number of data points, and p, the dimensionality of the data. The first type of data generating model we employ is what we call a "hypersphere-based approach". This model assumes the data is structured as a known number of hyper spherical masses. The probabilities of each mass, or component, are determined by a multinomial realization with equal probabilities. The locations of the centers of the masses are uniformly distributed over the support region (an arbitrarily scaled rectangle). For each mass, all points belonging to it are uniformly distributed on a hypersphere about the center. The radii of the hypersphere are all the same and should be scaled appropriately with the size of the support rectangle. This type of data generation falls in line most closely with the assumptions from k-means clustering. That is, that the data can be separated into a known number of distinct clusters, where the center of each cluster lies at the centroid of the points belonging to the cluster and all points belong to the cluster with the nearest centroid (Euclidean distance). This type of data is not exactly in line with k-means assumptions, as the k-means algorithm assumes that the data comes from a mixture of Gaussians with the restriction that all of the component Gaussians are spherical and have the same covariance structure. Rather than Gaussian, we let each hypersphere be uniform. This encourages distinct separation between clusters. This also allows hyperspheres to be overlapping at equal density, making the clustering process more challenging. The MOG clustering method is also fairly well suited for this type of data, as it can be thought of as a generalization of k-means without restricting covariance matrices to be the identity matrix (and allowing soft assignments to update parameters).

We also generate data is by using a mixture of Gaussians. The number of Gaussians in the mixture is deterministic. The probabilities of each mass, or component, are determined by a multinomial realization with equal probabilities. The means are uniformly distributed and the covariance matrices are randomly generated positive definite matrices constructed from columns of independent standard normal random variables. For each component, we randomly construct a positive definite covariance matrix by drawing the elements of a p dimensional matrix $\mathbf{Z}$ independently from $N(0,1)$ and taking

$\Sigma = \mathbf{Z}^T\mathbf{Z}$. The covariance matrices are scaled according to the size of the support region (a rectangle, as before) so there are chances of Gaussians overlapping at high densities. This type of data generation lends itself to using the EM algorithm to cluster using a mixture of Gaussians, as then the assumptions of this clustering algorithm are true in the data. To favor the constrained k-means constrained algorithm, we additionally generate data from a mixture of Gaussians with the constraint that an equal number of points are sampled from each component Gaussian.

The final data generating model we experiment with is constructed using a "trail-based approach". We generate trail- or density-based data to favor the DBSCAN algorithm. The data is generated in an iterative way. A trail is initialized with a single point, uniformly distributed on the support region. Then, we construct a trail by following a random walk, starting at the initial point. For each iterate in the random walk, there is a constant, specified probability that the random walk will end on that iterate. If the random walk ends, then the next iterate is a new random initialization to start a new random walk. If the random walk does not end, the iterate is obtained by taking a slight perturbation from the previous iterate. The perturbation is a realization of a multivariate normal distribution with identity covariance, scaled. Due to the random process, the number of "dense groups" in the sample will be unknown beforehand. However, we can set the "offspring probability" - that is, the probability of a random walk continuing in any iterate - such that the expected number of dense groups is known. This type of data model reasonably fits with density-based clustering. The cluster shapes are not necessarily convex; instead they are governed by local neighborhoods and the distance between points.

Finally, we use real financial data of daily returns for 49 assets during the period starting January 2013 and ending July 2015 available from the Center for Research in Security Prices. For this data, we consider the problem of portfolio optimization under conditional value-at-risk, rather than the newsvendor problem.

5.3.2 Simulations

All simulations were conducted in MATLAB. For each data generator, we compute average metric scores for each method over 500 replications, where a different random sample is generated in each replication and the clustering algorithms are applied to the

sample in the same way. The average is taken to get an estimate of expected performance of the algorithm. We repeat this process for different sets of parameter values to see how the number of points, number of clusters and dimension affect the results.

For k-means, we utilize the built in MATLAB function kmeans.m, which utilizes k-means++, which refers to the way the initial centers are constructed. We also set the replicates option to 5, so that five different initializations are performed and the best is taken (to improve the algorithm since the results change from initialization to initialization). For our constrained version of k-means, we implemented the alternating procedure of calculating cluster centers and solving the minimum-cost flow problem. For the mixture of Gaussians method, we employ the EM algorithm according to our own code in EM.m. This utilizes the iterative algorithm described in the methods section. For the DBSCAN method, we utilize code provided by Michal Daszykowski (2015). We set the minimum number of points equal to floor($\frac{N}{7k}$) and $\epsilon = .5$. As mentioned previously, we identify outliers as single points.

5.4 Results

The results are given in Tables 5.1 - 5.5. The most important results are in Table 5.1, which measures the average performance of RF over many trials after each clustering method has been used (as described in Section 5.2). The solution is computed for the newsvendor problem as described in Section 6.1. We are recording an estimate of the expected costs, given that the fragmentation is constructed through a given clustering technique. Thus, lower numbers indicate that the performance is better.

For most of the data types, constrained k-means performs the best, slightly edging out standard k-means. This is likely due to increased stability and smaller estimation errors, the dual motivations for designing the adaptation of k-means. Both standard and adapted EM are somewhat worse, and fairly comparable to each other. DBSCAN is in between the two pairs, except for the trail based data, in which it does best. This is intuitive as it is better able to capture the structure of the data in this case, even if the resulting clusters are nonconvex. For the real data, the EM methods do very well, and constrained k-means isn't as effective. This data isn't necessarily multimodal, so it seems that the simple mode identifying methods such as k-means rely strongly on the

assumption that the data is multimodal. Since RF is more useful on multimodal data, methods that perform better for such situations are of more interest, although RF can still be employed effectively on any data set. We look at the following metrics to have a better understanding of why these results occur.

In Table 5.2, the average number of clusters with a minimum eigenvalue below the threshold .01 is kept track of, as described in Section 5.2. A higher number indicates that on average more clusters are close to being degenerate, which will cause instability and estimation errors. The two types of k-means have the lowest scores for this metric for all types of data except the trail data, for which DBSCAN is the winner. Since k-means enforces convexity on the clusters, it is very unlikely to lead to extremely long, skinny groupings. Additionally, clusters are unlikely to be of very few points unless they are extreme outliers. Sparse clusters are much more compatible with the MOG and DBSCAN methods. Adapted k-means is even better than standard k-means because it discourages very small clusters, which lead to smaller eigenvalues. This is direct evidence that our constrained version of k-means is effective at increasing stability of the algorithm. All of the methods do significantly worse for the trail based data, since it is spread out and in nonconvex groupings, many of which may be overlapping. DBSCAN gains an advantage, likely because it results in fewer clusters of a very small number of points, since it is able to construct larger groupings through connectivity. Despite differences in the number of degenerate or almost degenerate clusters, there does not seem to be a significant difference in the computational time required for the different techniques, as given in Table 5.4. All of the methods are fairly comparable, with DBSCAN performing worse for most data types, except for the trail based data, once again. For the financial data, the k-means methods appear to have an edge over the MOG methods. This is likely because MOG takes longer to converge when clusters are not separated well or don't exist, and also may be due partially to the decreased stability due to the prevalence of clusters with small eigenvalues. Constrained k-means appears to have the best performance on data types for which it has the least number of degenerate clusters. It is likely that there is a correlation between good performance of RF, lack of sparse clusters, and low scores on the eigenvalue chart.

In Table 5.3, the within cluster sum-of-squares is recorded as described in Section 5.2. This shows how much information is lost through the data reduction. A lower value

indicates that less information is being thrown out. Since k-means directly minimizes this quantity, it has the lowest values. In general, constrained k-means has slightly higher WCSS due to the introduction of the penalty term into the objective criteria. The two EM methods follow, and DBSCAN has by far the highest WCSS in most cases. Keep in mind that this is multiplied by the number of clusters. Since DBSCAN may have any number of clusters, often it has more than the other methods and is therefore penalized. Additionally, clusters are not necessarily convex, leading to large sum-of-squares terms for individual clusters. Points may be assigned to a cluster by connectivity, when there is an alternative cluster whose center is actually much closer.

Finally, Table 5.5 contains the modified Davies-Bouldin Index we constructed in Section 5.2. This is an estimate of average overlap between regions in a fragmentation. Less overlap is considered better, as then adaptations to conditional moments estimation as described in Chapter 3 do not have to be made. We can see that k-means and constrained k-means are competitive for the hypersphere and equal MOG data types. However, for data where points are more sparsely spread out, such as MOG and trail data, k-means has much less overlap. The constrained version will require more points that are further away to be joined to clusters in order to achieve more equality in the population of each cluster. Both MOG methods are orders of magnitude higher in terms of overlap, since the regions are shaped like arbitrary ellipsoids. This may have a negative effect on the performance.

In summary, the constrained version of k-means is superior for both types of MOG and the hypersphere data. That is, all of the data types that have convex modes. This is probably due to the positive effects, in terms of estimation error and stability, of the evening out of cluster populations. Simple methods such as k-means and adaptations actually perform better than more complex models like MOG, due to factors such as less overlap between clusters and less information lost in reduction (WCSS). DBSCAN is not very effective except for trail-based data. Since this type of data is manufactured and the groupings are not modal, it doesn't provide strong evidence for using connectivity based methods for RF. For data that is not modal such as the real financial data, the MOG methods are able to outcompete k-means. This isn't as significant because RF is not as effective for this type of data in general.

Table 5.1: RF Performance

	$N = 40$, $p = 2$, $k = 4$, $\epsilon = .5$				
	k-means	Const. k-means	MOG	Adapted MOG	DBSCAN
Hypersphere	6.3420	6.2880	6.8638	7.0680	7.3797
MOG Equal	6.6363	5.8256	7.5972	7.2308	6.8876
MOG	7.1997	6.2909	7.7150	7.5977	7.1099
Trail	8.4762	8.4542	8.5644	8.4034	7.4132
Financial	0.0304	0.0453	0.0300	0.0199	0.0782

Table 5.2: Minimum Eigenvalues

	$N = 40$, $p = 2$, $k = 4$, $\epsilon = .5$				
	k-means	Const. k-means	MOG	Adapted MOG	DBSCAN
Hypersphere	0.0560	0.0800	0.1800	0.1520	0.6120
MOG Equal	0.3220	0.2620	0.4560	0.4740	0.6360
MOG	0.3520	0.2780	0.5520	0.5160	0.6820
Trail	1.4680	1.4560	1.6520	1.6400	1.2260
Financial	0.4420	0.3100	0.7400	0.6920	0.1080

Table 5.3: WCSS

	$N = 40$, $p = 2$, $k = 4$, $\epsilon = .5$				
	k-means	Const. k-means	MOG	Adapted MOG	DBSCAN
Hypersphere	67.5937	82.4027	88.7234	83.2142	567.3590
MOG Equal	340.1063	384.5172	437.4270	421.5246	654.2992
MOG	328.8587	373.6773	415.3481	407.6777	707.4011
Trail	58.3676	79.4875	77.0991	77.2406	89.3070
Financial	76.9440	90.7282	111.9205	107.9900	111.3271

Table 5.4: Computing Time

	$N = 40$, $p = 2$, $k = 4$, $\epsilon = .5$				
	k-means	Const. k-means	MOG	Adapted MOG	DBSCAN
Hypersphere	0.7176	0.7238	0.7148	0.7244	0.7283
MOG Equal	0.7239	0.7138	0.7193	0.7288	0.8151
MOG	0.7105	0.7011	0.7059	0.7173	0.8237
Trail	0.7569	0.7561	0.7710	0.7948	0.6831
Financial	0.8896	0.8446	0.9418	0.9591	0.8478

Table 5.5: Modified Davies-Bouldin Index

	$N = 40$, $p = 2$, $k = 4$, $\epsilon = .5$			
	k-means	Const. k-means	MOG	Adapted MOG
Hypersphere	0.2001	0.1215	18.6065	16.5306
MOG Equal	0.3216	0.3913	25.7285	31.6979
MOG	0.5347	0.7780	29.1846	17.4598
Trail	0.0331	1.6332	19.2496	25.5406

Chapter 6

Applications and Numerical Results

In this section, we will describe a selection of motivating applications and perform some numerical tests. All experiments were run on a Macintosh with a 2.4 GHz Intel processor and 8 GB memory. We use MATLAB R2014b and cvx modeling system to develop the code.

6.1 Newsvendor

The newsvendor model describes a classic mathematical framework for determining optimal inventory levels under uncertainty. Managers must make inventory decisions without knowing what the demand for a product(s) will be. The objective is to maximize profit or equivalently, minimize cost associated to overstocking and under stocking.

The newsvendor problem has been a test case for much of the work in DRO. Scarf's solution was derived for the newsvendor problem and included a unit-ordering cost in addition to holding cost and "stock-out" cost (1958). Zhu et al. (2006) consider limited distributional information for the newsvendor problem with ordering cost, but minimize a regret relative to the size of the cost, rather than a worst case cost. Gallego and Moon (1993) provide another proof of Scarf's ordering quantity when the mean and variance are known. They extend the results to include recourse, fixed ordering cost, and multiple items. Bertsimas and Thiel (2006) consider a stochastic linear program

where uncertainty is allowed in the constraint coefficients and total deviation of the coefficients from their nominal values is bounded by a set threshold. Perakis and Roels (2008) study the newsvendor problem with partial information, such as mean, variance, symmetry, etc. under a conservative but not robust setting. Yue et al. (2006) study the one dimensional newsvendor problem and focus on the expected value of distribution information (EVDI). There are also a variety of data-driven methods for the newsvendor model, such as in Godfrey and Powell (2001) and Levi et al. (2007).

Our setting for the newsvendor problem is as follows. There is a single seller, who must decide how many units $\mathbf{x}$ of product(s) to purchase in a given business cycle. Afterwards, customers will buy product quantities $\boldsymbol{\xi}$ and the seller will be charged a per-unit holding cost $\mathbf{h}$ for any leftovers that he/she is not able to sell and a per-unit backlog/loss of goodwill cost $\mathbf{b}$ for any unmet demand. We assume continuous variables so that any fraction of a unit may be sold. The cost is therefore $f(\mathbf{x}, \boldsymbol{\xi}) = \mathbf{b}^T(\boldsymbol{\xi} - \mathbf{x})^+ + \mathbf{h}^T(\mathbf{x} - \boldsymbol{\xi})^+$. We assume that a sample of historical demands is available.

6.1.1 Univariate

Consider the one dimensional newsvendor problem with equal backlog cost b and holding cost h. The SAA solution will be the sample median, $\hat{m}_N$. The MDRO solution will be the sample mean $\hat{\mu}_N$ (as shown in Scarf (1958)).

Now suppose the true distribution was $\mathcal{N}(\mu, \sigma^2)$, so that $\mu = m$. Both the SAA and MDRO estimators are unbiased. By the law of large numbers, $\hat{\mu}_N \to \mu = m$ and $\hat{m}_N \to m$ as $N \to \infty$, thus both estimators are consistent. By the Central Limit Theorem, $\hat{\mu}_N$ is asymptotically normal with $\text{Var}(\hat{\mu}_N) = \frac{\sigma^2}{N}$. On the other hand, as shown by Laplace, $\hat{m}_N$ is also asymptotically normal with $\text{Var}(\hat{m}_N) = \frac{\pi\sigma^2}{2N}$. Therefore, the MDRO solution is a more efficient estimator of the true solution asymptotically. This holds true for small N as well, as we can see in Figure 6.1. This example demonstrates how even an inconsistent or unbiased estimator, such as RF, may be superior to SAA for finite sample size, as we will see in numerical experiments.

Now consider the univariate newsvendor problem with arbitrary b and h. We consider an RF with two unbounded intervals and no probability ambiguity as follows:

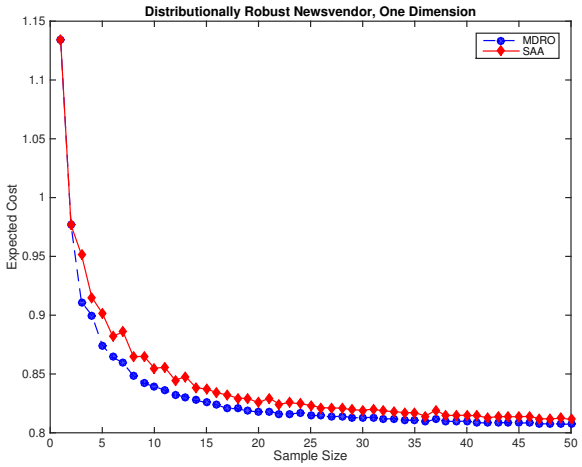

Figure 6.1: Efficiency of MDRO and SAA estimators

$$
\begin{aligned}
\min_{\mathbf{x}} \quad \sup_{\pi\in\Theta} \quad & \mathbb{E}_\pi[b(\Xi - x)^+ + h(x-\Xi)^+], \\
\text{subject to} \quad & \mathbb{P}_\pi[\Xi \le \beta] = \hat{p}_1, \\
& \mathbb{P}_\pi[\Xi \ge \beta] = \hat{p}_2, \\
& \mathbb{E}_\pi[\Xi | \Xi \le \beta] = \hat{\mu}_1, \\
& \mathbb{E}_\pi[\Xi | \Xi \ge \beta] = \hat{\mu}_2, \\
& \mathbb{E}_\pi[\Xi^2 | \Xi \le \beta] = \hat{\sigma}_1^2 + \hat{\mu}_1^2, \\
& \mathbb{E}_\pi[\Xi^2 | \Xi \ge \beta] = \hat{\sigma}_2^2 + \hat{\mu}_2^2,
\end{aligned}
\tag{6.1}
$$

where π lies in the space of probability measures Θ over the measurable space $(\mathbb{R}, \mathscr{F})$.

Theorem 7. *The solution to* (6.1) *is given by:*

$$
x_\star = \left\{ \begin{array}{ll}
\hat{\mu}_1 + \frac{\bar{r}_2\sigma_1}{\sqrt{1-\bar{r}_2^2}}, & \textit{if (i)} \\
\beta - \frac{(\beta-\hat{\mu}_1)^2+\hat{\sigma}_1^2}{2(\beta-\hat{\mu}_1)}, & \textit{if (ii)} \\
\beta, & \textit{if (iii)} \\
\beta + \frac{(\hat{\mu}_2-\beta)^2+\hat{\sigma}_2^2}{2(\hat{\mu}_2-\beta)}, & \textit{if (iv)} \\
\hat{\mu}_2 - \frac{\bar{r}_1\sigma_2}{\sqrt{1-\bar{r}_1^2}}, & \textit{if (v)}
\end{array} \right\},
$$

Conditions (i)-(v) are outlined in the proof.

Proof. The proof is given in Appendix B.4. □

6.1.2 Multivariate

Now we consider the general multivariate newsvendor. Since f is convex and piece-wise linear in decision variable $\mathbf{x}$ and $\boldsymbol{\xi}$, it satisfies Assumption 1. The problem with probability ambiguity will be:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{\gamma_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \eta+\rho\lambda+\lambda\sum_{k=1}^{r}\hat{p}_k\phi^\star\left(\frac{z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta}{\lambda}\right)\\
& +\lambda\sum_{k=r+1}^{K}\hat{p}_k\phi^\star\left(\frac{\gamma_k-\eta}{\lambda}\right),\\
\text{subject to} \quad & \begin{pmatrix}1 & \boldsymbol{\xi}^T\end{pmatrix}\begin{pmatrix}z_k+(\mathbf{a}_l\circ(\mathbf{h}+\mathbf{b}))^T\mathbf{x} & \frac{1}{2}(\mathbf{z}_k-(\mathbf{a}_l\circ(\mathbf{h}+\mathbf{b})))^T\\ \frac{1}{2}(\mathbf{z}_k-(\mathbf{a}_l\circ(\mathbf{h}+\mathbf{b}))) & \mathbf{Z}_k\end{pmatrix}\begin{pmatrix}1\\ \boldsymbol{\xi}\end{pmatrix}\\
& \geq 0,\forall\boldsymbol{\xi}\in\Omega_k,\forall l=1,...,2^p,\forall k=1,...,r,\\
& z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\forall k=1,...,r,\\
& \gamma_k\geq\ -(\mathbf{a}_l\circ(\mathbf{h}+\mathbf{b}))^T\mathbf{x}+(\mathbf{a}_l\circ(\mathbf{h}+\mathbf{b}))^T\mathbf{m}_k,\ \forall\ l=1,...,2^p,\\
& \forall k=r+1,...,K,\\
& \gamma_k-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\ \forall k=r+1,...,K,\\
& \lambda\geq 0.
\end{aligned}
\tag{6.2}
$$

where p is the dimension of $\mathbf{x}$, $\circ$ denotes element-wise multiplication, and $\mathbf{a}_l$ is a vector of ones and zeroes for each l, where $l=1,...,2^p$ indexes through all such unique vector combinations.

In the following experiments, we solve the original problem without probability

ambiguity, which is:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\{\gamma_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \sum_{k=1}^{r}\hat{p}_k\left(z_k+\mathbf{z}_k{}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)\right)+\sum_{k=r+1}^{K}\hat{p}_k\gamma_k,\\
\text{subject to} \quad & \begin{pmatrix}1 & \boldsymbol{\xi}^T\end{pmatrix}\begin{pmatrix}z_k+(\mathbf{a}_l\circ(\mathbf{h}+\mathbf{b}))^T\mathbf{x} & \frac{1}{2}(\mathbf{z}_k-(\mathbf{a}_l\circ(\mathbf{h}+\mathbf{b})))^T\\ \frac{1}{2}(\mathbf{z}_k-(\mathbf{a}_l\circ(\mathbf{h}+\mathbf{b}))) & \mathbf{Z}_k\end{pmatrix}\begin{pmatrix}1\\ \boldsymbol{\xi}\end{pmatrix}\geq 0,\\
& \forall\boldsymbol{\xi}\in\Omega_k,\forall l=1,...,2^p,\ \forall k=1,...,r,\\
& \gamma_k\geq\ -(\mathbf{a}_l\circ(\mathbf{h}+\mathbf{b}))^T\mathbf{x}+(\mathbf{a}_l\circ(\mathbf{h}+\mathbf{b}))^T\mathbf{m}_k,\ \forall\ l=1,...,2^p,\\
& \forall k=r+1,...,K.
\end{aligned}
\tag{6.3}
$$

As we note, regardless of the fragmentation regions Ω_k, the formulation is NP hard since the number of constraints grow exponentially in p. This is even true for MDRO and it is also a copositive program as $\boldsymbol{\xi}\geq 0$. If we fragment into ellipsoids and differences of ellipsoids as in (2.10) for an ellipsoidal partition, the problem remains NP hard due to size, which is unavoidable, but becomes an SDP (with no probability ambiguity) and thus easier than the MDRO version. With probability ambiguity, it is an SCP. In any case, we could use quadratic decision rules to get an upper bound as in Hanasusanto et al. (2014). We do not and directly solve the original problem since we are dealing with $p=1$ or 2 (with no probability ambiguity).

For our numerical experiments, we consider the two dimensional newsvendor problem without probability ambiguity, which it is still meaningful for practical applications. For example, suppose that the seller is selling two types of clothing items. The items can go together, so that if one is in fashion during the business cycle, the other is more likely to be as well. Both are more likely to be out of fashion than one in and one out. Thus, the distribution for the demand of the two products has several modes depending on fashion trends which are unknown to the seller prior to their decision.The procedure for the experiment is as follows. We generate N data points from a mixture of Gaussians. For each method, we estimate parameters and formulate the optimization problem as in (6.2), the solution of which gives us the optimal inventory levels $\mathbf{x}^\star$. For RF, we set $r=4$. In all cases we compute out-of-sample costs $\mathbf{b}^T(\mathbf{d}-\mathbf{x}^\star)^+ + \mathbf{h}^T(\mathbf{x}^\star-\mathbf{d})^+$, where

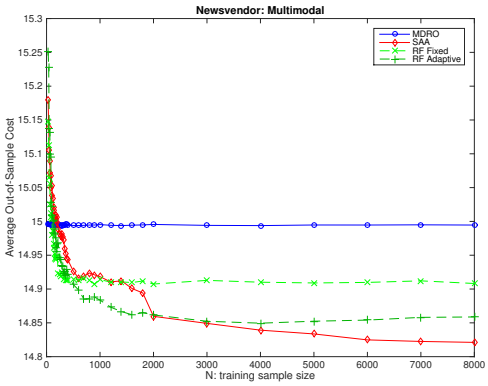

Figure 6.2: Relative RF Performance on NV: Large N

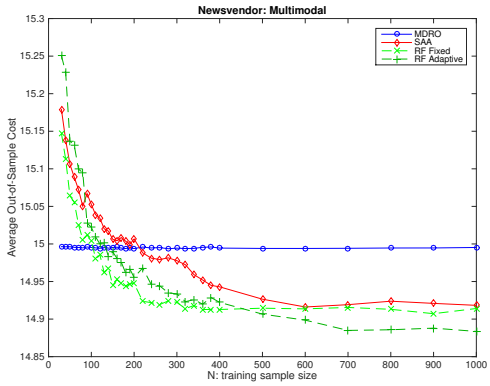

Figure 6.3: Relative RF Performance on NV: Small N

$\mathbf{h} = \mathbf{b} = \mathbf{e}$, on a sample of $N = 80000$ to get an estimate of the expected cost. We repeat this procedure 400 times and consider average expected costs over the experiments.

Figure 6.2 and Figure 6.3 illustrate performance of RF in relation to MDRO and SAA. Both plots contain the same data; the plot on the right zooms in on smaller N values from the plot on the left. "RF Fixed" refers to an implementation where the regions are fixed for all samples using a priori knowledge, where the regional ellipsoids are conservatively large as to contain all sample points with high probability. "RF Adaptive" represents the implementation we describe in Chapter 4 where each region is constructed by clustering and minimum volume containing ellipsoids. We observe that one of the two RF methods perform best for intermediate values of N. As N grows, the SAA solution relatively improves, but we are not adjusting RF to the sample size (r is fixed at 4). For smaller N, RF performs better because it is a simpler model (fewer parameters). The fixed RF approach is superior for small N due to its increased stability. As N increases, the adaptive approach outstrips since it fits directly to the sample and the ellipsoids will be smaller. For very large N, SAA may perform slightly better but be more computationally expensive. The SAA problem size scales with N, while RF has fixed problem size, as evident in the computational time required to solve the optimization problem in cvx as demonstrated in Figure 6.4 (RF here is adaptive).

In Figure 6.5, we demonstrate performance of RF (the adaptive approach) when the

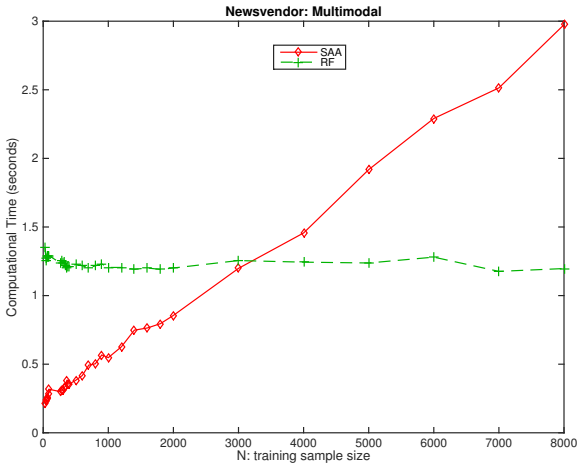

Figure 6.4: Computational Time

generating distribution is changing over time. The experiment is similar to that illustrated in Figures 6.2 and 6.3. The initial generating distribution is the same multimodal distribution in $\mathbb{R}^2$. However, on each day, the mean vector is perturbed (uniformly) and the new data is generated from the perturbed distribution. For each method, we use the past N days to compute moments and generate a solution, whose performance is measured on the current day. This is repeated over 100 days. The full experiment is then repeated 50 times for each value of N and performance is averaged. Since the perturbations may move the mean vectors of each mode further apart or closer together in each experiment, there is more variability than with a constant distribution. Additionally, as N increases, more accurate estimates of the past moments are not necessarily useful, as the distribution is changing. After N reaches about 300, larger sample size does not seem to provide a benefit. It may even be harmful as the perturbations accumulate and a larger sample size is slower to adjust.

It is clear that RF performs the best under the conditions of a changing distribution for these values of N. It doesn't make sense to consider larger values of N, since data too far in the past is not relevant. For very small sample size, MDRO is superior, as before, but as N passes 100, RF and SAA surpass it. RF is slightly better than SAA for almost every value of N. This may be due to less variability. RF appears to be more robust than SAA when the distribution is changing over time.

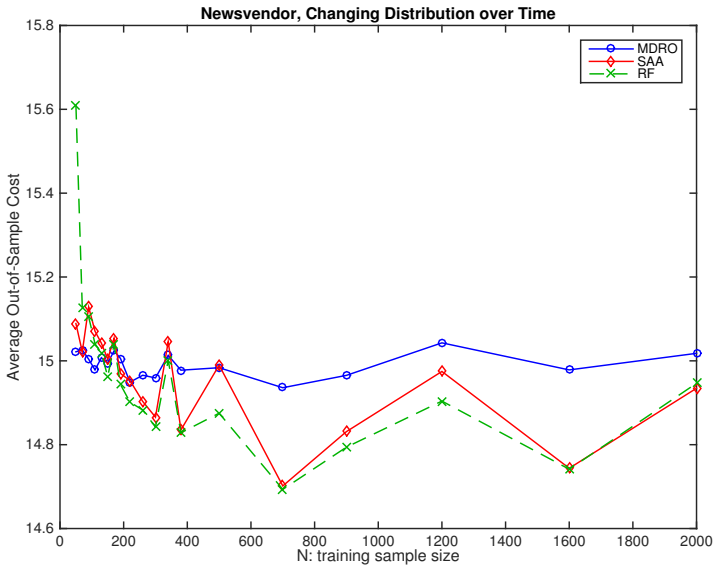

Figure 6.5: Performance when Distribution Changes Over Time

In Figure 6.7, we demonstrate how RF, without probability ambiguity, can be thought of as a bridge between MDRO and SAA. We generate samples of size $N = 120$ from a piecewise uniform distribution over a rectangle in $\mathbb{R}^2$. We fragment into a polytopal partition of rectangles of equal size for different values of r. This is depicted in Figure 6.6, where the green points are an example sample and $r = 64$. As the number of regions increases, the average number of points in each region decreases. We plot the objective value, the worst case expected cost for the newsvendor problem, averaged over many samples of size $N = 120$, on the right. With one region, RF is closest to MDRO (it is not equivalent since the domain is bounded). For a given sample, as r increases, the objective value must decrease as the ambiguity set is further restricted. As r increases, more regions will contain only one point and, as $r \to \infty$, the objective value will converge to that of the SAA.

6.2 Facility Location

Suppose that we wish to determine the optimal location for a single facility. For example, we need to locate a store or a hospital. Our decision variable $\mathbf{x} \in \mathbb{R}^2$ is the chosen

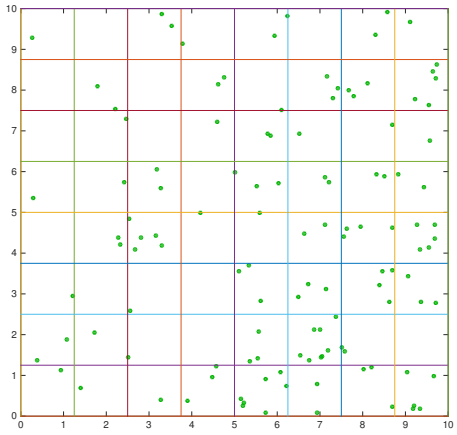

Figure 6.6: Polytopal Partition

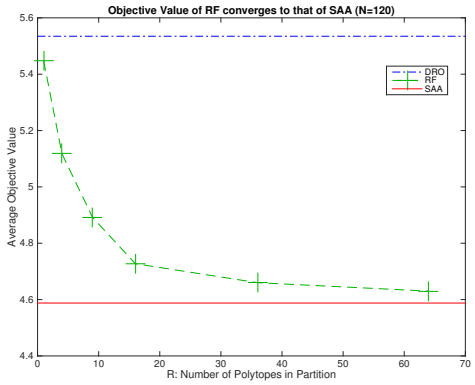

Figure 6.7: Objective Value of RF and SAA

location of the facility. We represent the demand of our customers $\boldsymbol{\xi}$ as a random vector in $\mathbb{R}^2$. If we wish to minimize distance traveled for our customers or the community to the facility then our cost function $f(\mathbf{x}, \boldsymbol{\xi}) = \|\mathbf{x} - \boldsymbol{\xi}\|$ where $\|\cdot\|$ represents some measure of distance. If we consider the L_1 norm, the cost function is piecewise linear since $\|\mathbf{x} - \boldsymbol{\xi}\|_1 = |x_1 - \xi_1| + |x_2 - \xi_2|$. Since it is naturally restricted to two dimensions, we could use either a quadrilateral or an ellipsoidal fragmentation and get a tractable formulation. This application demonstrates that the formulation in (2.8) is valuable for some applications. Robust fragmentation will enable use of regional customer information. This will be especially valuable if the demand distribution is organized into clusters based on population density, geography, and/or other factors.

We also consider the L_2 norm. In this case f does not satisfy Assumption 2, and so we cannot apply Theorem 1. However, we present the MDRO version of it as it requires a different technique and provides a nice demonstration of how the cost function may

be simply linear or quadratic in some applications. Consider:

$$\begin{aligned} \min_{\mathbf{x}\in\mathscr{X}} \quad \sup_{\pi\in\Theta} \quad & \mathbb{E}_\pi[\|\mathbf{x}-\Xi\|_2], \\ \text{subject to} \quad & \mathbb{E}_\pi[\Xi] = \hat{\boldsymbol{\mu}}, \\ & \mathbb{E}_\pi[\Xi\Xi^T] = \hat{\boldsymbol{\Sigma}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T, \\ & \mathbb{P}_\pi[\Xi \in \mathbb{R}^p] = 1. \end{aligned} \tag{6.4}$$

Theorem 8. *The solution to problem* (6.4) *is given by the following semidefinite program:*

$$\begin{aligned} \min_{\mathbf{x}\in\mathscr{X},z,\mathbf{z},\mathbf{Z}} \quad & z + \mathbf{z}^T\hat{\boldsymbol{\mu}} + \mathbf{Z}\circ(\hat{\boldsymbol{\Sigma}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T), \\ \textit{subject to} \quad & \begin{pmatrix} z-\tau & -\frac{1}{2}\mathbf{x}^T & \frac{1}{2}\mathbf{z}^T \\ -\frac{1}{2}\mathbf{x} & \tau\mathbf{I}_p & \frac{1}{2}\mathbf{I}_p \\ \frac{1}{2}\mathbf{z} & \frac{1}{2}\mathbf{I}_p & \mathbf{Z} \end{pmatrix} \succeq 0, \\ & \tau \geq 0, \end{aligned}$$

where $\mathbf{I}_p$ *is the* p*-dimensional identity matrix.*

Proof. Observe that $\|\mathbf{x}-\boldsymbol{\xi}\|_2 = \max_{\|\boldsymbol{\eta}\|_2\leq 1} \ \boldsymbol{\eta}^T(\mathbf{x}-\boldsymbol{\xi})$. Therefore, the constraint is equivalent to $\boldsymbol{\xi}^T\mathbf{Z}\boldsymbol{\xi}+\mathbf{z}^T\boldsymbol{\xi}+z \geq \boldsymbol{\eta}^T(\mathbf{x}-\boldsymbol{\xi}), \ \ \forall\boldsymbol{\xi}\in\mathbb{R}^p, \forall\|\boldsymbol{\eta}\|_2^2 \leq 1$. Thus, we are requiring nonnegativity of a quadratic function in $\begin{pmatrix}\boldsymbol{\xi}\\ \boldsymbol{\eta}\end{pmatrix}$ over a region described by a quadratic inequality, $\|\boldsymbol{\eta}\|_2^2 \leq 1$. Therefore, we can use S-Lemma to convert this to the following constraint: $\boldsymbol{\xi}^T\mathbf{Z}\boldsymbol{\xi} + \mathbf{z}^T\boldsymbol{\xi} + z - \boldsymbol{\eta}^T(\mathbf{x}-\boldsymbol{\xi}) - \tau(1-\|\boldsymbol{\eta}\|_2^2) \geq 0 \ \ \forall\boldsymbol{\xi},\boldsymbol{\eta}\in\mathbb{R}^p$, where $\tau \geq 0$. Representing this as an LMI, we get the formulation in Theorem (8). □

Some other authors have considered versions of robust facility location, although different from above. Wu et al. (2015) consider a two stage problem where locations are fixed. Their problem is to determine which stage to open facilities in given random demand and first and second moments constraints. They extend LP approximation algorithms from non DRO to DRO with the same approximation ratios. Bertsimas et al. (2010) investigate minimax two-stage stochastic linear optimization formulations under first and second moment constraints. The methods and results are similar to Wu

et al. (2015); they apply the concept to production-transportation and facility location, demonstrating that performance is better than data-driven methods under extremal distributions.

6.3 Portfolio Optimization under Conditional Value-at-Risk

A common approach to optimizing a portfolio of assets $\mathbf{x}$ to limit risk is to consider the conditional value-at-risk (CVaR). CVaR, also known as mean excess loss, has been shown to be an effective measure of risk. We consider a loss function $g(\mathbf{x}, \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ is the uncertain return. For example, we may let $g(\mathbf{x}, \boldsymbol{\xi}) = -\boldsymbol{\xi}^T\mathbf{x}$, the negative of the return to the investor. The objective is then to minimize CVaR under the loss function. The probability of the loss not exceeding a threshold γ is given by $\Psi(\mathbf{x}, \gamma) = \int_{g(\mathbf{x},\boldsymbol{\xi})\leq\gamma} 1 d\pi(\boldsymbol{\xi})$. Thus, the α-VaR is given by $\gamma_\alpha(\mathbf{x}) = \min\{\gamma : \Psi(\mathbf{x}, \gamma) \geq \alpha\}$. The CVaR is given by the expected loss given that the loss does exceed the α threshold, $\phi_\alpha(\mathbf{x}) = \frac{1}{1-\alpha}\int_{g(\mathbf{x},\boldsymbol{\xi})\geq\gamma_\alpha(\mathbf{x})} g(\mathbf{x}, \boldsymbol{\xi}) d\pi(\boldsymbol{\xi})$. An equivalent formulation for CVaR is $\phi_\alpha(\mathbf{x}) = \min_\gamma\{\gamma + \frac{1}{1-\alpha}\int_{\boldsymbol{\xi}}[g(\mathbf{x}, \boldsymbol{\xi}) - \gamma]^+ d\pi(\boldsymbol{\xi})\}$ (for details, see Rockafellar and Uryasev (2000)). If we let $f(\gamma, \mathbf{x}, \boldsymbol{\xi}) = \gamma + \frac{1}{1-\alpha}[g(\mathbf{x}, \boldsymbol{\xi}) - \gamma]^+$, then we see that the objective can be written as $\min_{\gamma,\mathbf{x}} \mathbb{E}_\pi[f(\gamma, \mathbf{x}, \Xi)]$. We can formulate the robust version as in (2.2) and solve:

$$\min_{\mathbf{x},\gamma} \quad \sup_{\pi\in\Theta} \left[\int_{\boldsymbol{\xi}} (\gamma + \frac{1}{1-\alpha}[g(\mathbf{x}, \boldsymbol{\xi}) - \gamma]^+) d\pi(\boldsymbol{\xi})\right], \tag{6.5}$$

with ambiguity set Θ. The cost function f satisfies Assumption 1 if g does. Typically we consider $g(\mathbf{x}, \boldsymbol{\xi}) = -\boldsymbol{\xi}^T\mathbf{x}$, so that $m = 2$ from (2.4) regardless of p. Thus, the problem size is polynomial in the number of regions r. Additionally, we generally have resource constraints given by $\mathbf{e}^T\mathbf{x} \leq 1$ and $\mathbf{x} \geq 0$. A constraint on the expected return is often enforced as well, for example, $\mathbf{x}^T\boldsymbol{\mu} \geq R$ where R is a required minimum.

We consider the following setup for the portfolio selection problem. A similar setup has been considered by Delage and Ye (2010) and Wang et al. (2013). We gather historical data of daily returns for 49 assets during the period starting January 2013 and ending July 2015 available from the Center for Research in Security Prices. We performed 30 experiments (each with different assets, selected randomly) and averaged over them. In each experiment, we select four assets and the decision is to select a

portfolio from these four assets for each day in the time period. The objective is to make a portfolio allocation $\mathbf{x}$ to minimize the conditional value-at-risk, where $g(\mathbf{x}, \boldsymbol{\xi}) = -\boldsymbol{\xi}^T\mathbf{x}$. We also keep track of the daily returns. On a given day, the decision maker observes a sample of $N = 60$ or $N = 120$ realizations of the most recent daily returns $\boldsymbol{\xi}_j^i$, for assets $j = 1, ..., p$, and days $i = 1, ..., N$. Given the sample, we compare three different methods. The SAA model minimizes (6.5) where the integral is replaced with the sum over the sample. The MDRO model minimizes (6.5) where the ambiguity set Θ is constructed from the estimated mean and covariance of the sample. For the RF approach, we use a fixed $r = 4$, and cluster according to constrained k-means as described in the previous section. Then we solve (6.5) where the Θ is constructed according to the conditional moments estimates. For simplicity, we do not construct an ambiguity set for the probability vector. We continue this procedure for every day over the considered time period.

The results are given in Tables 6.1 and 6.2. The out-of-sample CVaR is the negative of the average return out of the $1 - \alpha$ worst daily returns in each experiment, where $\alpha = .95$, and averaged over all experiments. The mean return is computed over all days for each experiment and then averaged over experiments. The third column gives the variance in the daily returns, averaged over all experiments. For $N = 60$, RF clearly performs best in terms of the objective, conditional value-at-risk. RF also has a much lower variance in the daily returns. Here the variance is computed over the time period and averaged over experiments. RF does have the lowest mean return. SAA results in the worst CVaR and the highest variance in returns, but also the highest mean return. As the objective is to minimize CVaR, RF is performing very well. However, in order to minimize exposure to the worst case returns, there must be some sacrifice in terms of the overall mean return.

For $N = 120$, MDRO outperforms RF in terms of CVaR. However, RF still has the lowest variance in daily returns, a desirable property. Additionally, RF has the highest daily mean return. It is interesting to note that with larger sample size, the CVaR decreases substantially, but the daily mean returns actually decrease. Performance improves in terms of avoiding risk, but this comes at the cost of reduced daily returns. Some of this effect may also be the product of different experiments run for the two sample sizes (assets are not guaranteed to be the same).

Table 6.1: Portfolio Optimization under Conditional Value-at-Risk

	$N = 60$		
	Out-of-Sample CVaR	**Mean Return**	**Variance of Returns**
RF	1.0890	0.0258	0.2114
SAA	1.1549	0.0284	0.2601
MDRO	1.1321	0.0279	0.2507

Table 6.2: Different Parameter Values

	$N = 120$		
	Out-of-Sample CVaR	**Mean Return**	**Variance of Returns**
RF	0.3524	0.0111	0.0756
SAA	0.3556	0.0107	0.0787
MDRO	0.3509	0.0107	0.0762

We conclude that RF performs comparatively well when employed in portfolio allocation. It results in low variance in returns, a desirable feature, and competitive mean daily returns and conditional value-at-risk. It is effective at minimizing risk in comparison to SAA when measured by CVaR, especially for smaller N.

Chapter 7

Black Swan Events

The phrase "black swan" terms from old Latin, hearkening back to use by the poet Juvenal when describing an impossible event. Black swans were unknown in Europe and assumed not to exist. All historical records showed that swans were white. After a Dutch explorer discovered black swans in Australia in the late seventeenth century and brought the news back to England, the phrase transformed to mean a supposed impossibility that is later proven to be possible.

Nassim Nicholas Taleb took the idea further by describing black swan events in his books *Fooled By Randomness* and *The Black Swan*. These events are unforeseen, unpredictable, and extremely high profile. Taleb began by using black swan events to explain financial markets, but then extended the concept to all kinds of events, eventually arguing that almost all historical events of importance could be classified in this way. A black swan event has three main characteristics. First, it is an unforeseen outlier, as it has never been observed before. Second, it is very impactful. Third, after its occurrence, it is rationalized and explained due to a need to give meaning and order to events in our world. (Taleb (2007))

Due to their very nature, black swan events are unpredictable. Thus, the only way to combat negative black swans is to design systems, or make decisions, that are robust against them. Taleb contends that the difficulty lies in the paucity of similar events in the historical record. When the probabilities are small but consequences are large, a priori assumptions may have to be made to account for black swans. In reality, events have distributions with much fatter tails than assumed and the important ones are the

shocks and large deviations.

Robust fragmentation, although more robust than SAA, is still vulnerable to black swan events. This is because it uses a historical sample to construct the probabilities and support of distributions in the ambiguity set. RF does not account for events that have never occurred. While we are not making strong assumptions like normality, which Taleb calls the "great intellectual fraud", we are still relying on observations to guide our decision making. In order to protect against rare and unforeseen events, we must allow for the possibility of events outside the realm of what we have seen. We can fortify our method against black swan events by loosening our ambiguity set to include more of the unpredictable. In this chapter, we will discuss how we can account for black swan events in robust fragmentation.

7.1 Ambiguity in Support

We account for unpredictable, rare events by extending the support of the distributions in our ambiguity set to cover areas outside the historical data sample. After we have used the data to obtain a fragmentation, we construct an additional region Ω_0, which is a difference of ellipsoids outside the current support. This region is intended to cover events with a large magnitude (far from the center of the data sample), but with low probability (outside the observed support). The inner ellipsoid is generated as the minimum volume containing ellipsoid for the data and the outer ellipsoid is given the same center and shape, with a radius determined as $S\delta_0$, where δ_0 is the inner radius. Figures 7.1 and 7.2 illustrate an example of this. Figure 7.1 gives a bird's eye view of the data, the fragmentation, and the black swan region (between the outer two ellipsoids). Figure 7.2 zooms in on the data sample to give a clearer idea of the regions.

As can be seen, we are accounting for events that have not even occured in the historical data in order to provide robustness to them. We then assign a probability p_0 to this region and rescale the probability estimates of all the other regions $\tilde{p}_k = \hat{p}_k(1-p_0)$ so that they are consistent. The probability p_0 should be large enough to provide robustness, but not so large that the method becomes overly conservative. We choose a value of $p_0 = .01$ in our experiments. This parameter should also be balanced with the ratio S of the outer radius to the inner radius in the black swan region. The larger this

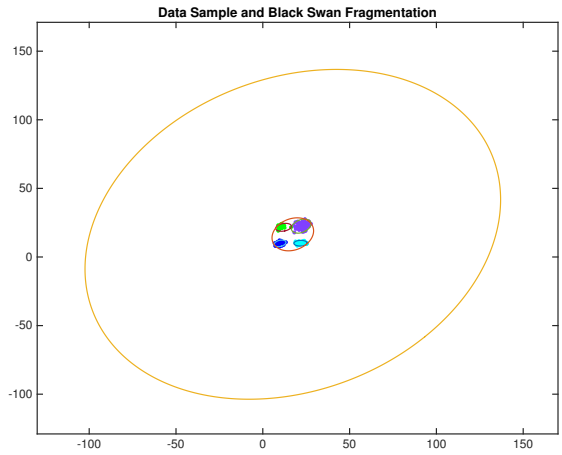

Figure 7.1: Data Sample and Black Swan Region, Zoomed Out

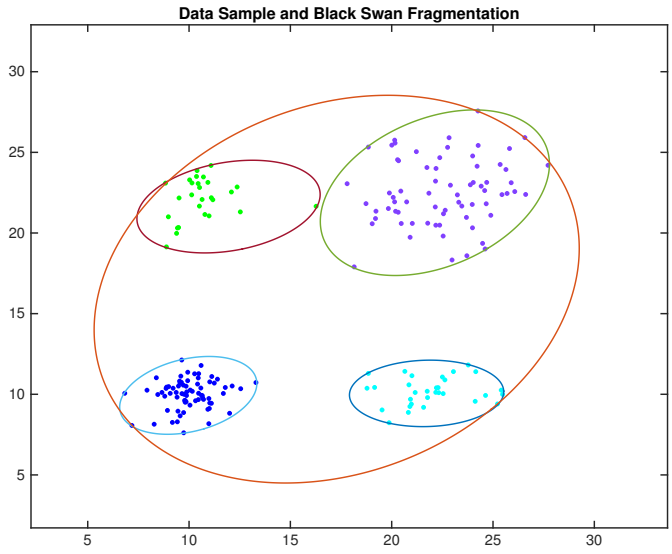

Figure 7.2: Data Sample and Black Swan Region, Zoomed In

ratio, the more extreme events are considered, which should be assigned an even lower probability. Let's return to the RF formulation as in (2.3) with this additional region:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X}} \quad & \sup_{\mathbf{p}\in\mathscr{P}} \sum_{k\in C} p_k \sup_{\pi_k\in\Theta_k} \int_{\Omega_k} f(\mathbf{x},\boldsymbol{\xi}) d\pi_k(\boldsymbol{\xi}) + \sum_{k\in S} p_k f(\mathbf{x},\mathbf{m}_k) \\
& \quad + p_0 \sup_{\pi_0\in\Theta_0} \int_{\Omega_0} f(\mathbf{x},\boldsymbol{\xi}) d\pi_0(\boldsymbol{\xi}), \\
\text{subject to} \quad & \mathbf{e}^T\mathbf{p} = 1 - p_0, \\
& \mathbf{p} \geq 0, \\
& \sum_{k=1}^{K} \tilde{p}_k \phi(\frac{p_k}{\tilde{p}_k}) \leq \rho, \\
& \int_{\Omega_k} d\pi_k(\boldsymbol{\xi}) = 1, \\
& \int_{\Omega_k} \boldsymbol{\xi} d\pi_k(\boldsymbol{\xi}) = \hat{\boldsymbol{\mu}}_k, \\
& \int_{\Omega_k} \boldsymbol{\xi}\boldsymbol{\xi}^T d\pi_k(\boldsymbol{\xi}) = \hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T, \ \forall k = 1, ..., r, \\
& \int_{\Omega_0} d\pi_0(\boldsymbol{\xi}) = 1,
\end{aligned}
\tag{7.1}
$$

where π_0 is the conditional distribution for $\boldsymbol{\xi}$ given that it is in Ω_0. Note that there are no moments constraints for the additional region Ω_0, and that $\mathbf{p}$ does not include p_0.

Theorem 9. *Problem* (7.1)*, when Assumption 2 holds, can be reformulated as:*

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{\gamma_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & (1-p_0)\eta + \rho\lambda + \lambda\sum_{k=r+1}^{K}\tilde{p}_k\phi^\star\left(\frac{\gamma_k-\eta}{\lambda}\right) + \lambda p_0\phi^\star\left(\frac{z_0-\eta}{\lambda}\right), \\
& + \lambda\sum_{k=1}^{r}\tilde{p}_k\phi^\star\left(\frac{z_k + \mathbf{z}_k{}^T\hat{\boldsymbol{\mu}}_k + \mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta}{\lambda}\right), \\
\textit{subject to} \quad & \begin{pmatrix}1 & \boldsymbol{\xi}^T\end{pmatrix}\begin{pmatrix} z_k - c_l - \boldsymbol{\beta}_l^T\mathbf{x} & \frac{1}{2}(\mathbf{z}_k - \mathbf{b}_l - \mathbf{D}_l^T\mathbf{x})^T \\ \frac{1}{2}(\mathbf{z}_k - \mathbf{b}_l - \mathbf{D}_l^T\mathbf{x}) & \mathbf{Z}_k - \mathbf{A}_l\end{pmatrix}\begin{pmatrix}1\\ \boldsymbol{\xi}\end{pmatrix} \geq 0 \\
& \forall\boldsymbol{\xi}\in\Omega_k, \forall l = 1,...,L,\ \forall k = 1,...,r, \\
& z_k + \mathbf{z}_k{}^T\hat{\boldsymbol{\mu}}_k + \mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T) - \eta \leq \left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda, \\
& \forall k = 1,...,r, \\
& \gamma_k \geq (\boldsymbol{\beta}_l + \mathbf{D}_l\mathbf{m}_k)^T\mathbf{x} + \mathbf{m}_k^T\mathbf{A}_l\mathbf{m}_k + \mathbf{b}_l^T\mathbf{m}_k + c_l,\ \forall\, l = 1,...,L, \\
& \forall k = r+1,...,K, \\
& \gamma_k - \eta \leq \left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\ \forall k = r+1,...,K, \\
& \lambda \geq 0, \\
& \begin{pmatrix}1 & \boldsymbol{\xi}^T\end{pmatrix}\begin{pmatrix} z_0 - c_l - \boldsymbol{\beta}_l^T\mathbf{x} & \frac{1}{2}(-\mathbf{b}_l - \mathbf{D}_l^T\mathbf{x})^T \\ \frac{1}{2}(-\mathbf{b}_l - \mathbf{D}_l^T\mathbf{x}) & -\mathbf{A}_l\end{pmatrix}\begin{pmatrix}1\\ \boldsymbol{\xi}\end{pmatrix} \geq 0\ \forall\boldsymbol{\xi}\in\Omega_0, \\
& \forall l = 1,...,L, \\
& z_0 - \eta \leq \left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda.
\end{aligned}
\tag{7.2}
$$

Proof. We compute the dual of the problem exactly as before in the proof of Theorem 1, except that we have the one additional supremum term. We associate dual variable z_0 with the constraint $\int_{\Omega_0} d\pi_0(\boldsymbol{\xi}) = 1$, thus we have Lagrangian $z_0 + \int_{\Omega_0}[f(\mathbf{x},\boldsymbol{\xi}) - z_0]d\pi_0(\boldsymbol{\xi})$ for the final supremum term. This is bounded above if and only if $f(\mathbf{x},\boldsymbol{\xi}) - z_0 \leq 0$ for all $\boldsymbol{\xi}\in\Omega_0$. In this case, the maximum occurs at $\pi_0 = 0$ everywhere, thus we have a slightly modified version of (2.4): The only additions are an additional term in the objective function, which follows the same form as the others, and two new sets of constraints. The constraint $z_0 - \eta \leq \left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda$ is just a linear constraint. The new infinite

dimensional constraints, $\begin{pmatrix}1 & \boldsymbol{\xi}^T\end{pmatrix}\begin{pmatrix} z_0 - c_l - \boldsymbol{\beta}_l^T\mathbf{x} & \frac{1}{2}(-\mathbf{b}_l - \mathbf{D}_l^T\mathbf{x})^T \\ \frac{1}{2}(-\mathbf{b}_l - \mathbf{D}_l^T\mathbf{x}) & -\mathbf{A}_l \end{pmatrix}\begin{pmatrix}1 \\ \boldsymbol{\xi}\end{pmatrix} \geq 0 \ \forall \boldsymbol{\xi} \in \Omega_0$, are, once again, nonnegativity of a quadratic function in $\boldsymbol{\xi}$, over a region Ω_0. This is why we formulate the outer region as a difference of ellipsoids. Then, as we did in Theorem 3, we can use Lemma 2 to rewrite the constraint as an LMI:
$\begin{pmatrix} z_0 - c_l - \boldsymbol{\beta}_l^T\mathbf{x} & \frac{1}{2}(-\mathbf{b}_l - \mathbf{D}_l^T\mathbf{x})^T \\ \frac{1}{2}(-\mathbf{b}_l - \mathbf{D}_l^T\mathbf{x}) & -\mathbf{A}_l \end{pmatrix} - \tau_{0_1}\begin{pmatrix} -\delta_0^2 + \boldsymbol{\nu}_0^T\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\nu}_0 & -(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\nu}_0)^T \\ -\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\nu}_0 & \boldsymbol{\Lambda}_0^{-1} \end{pmatrix}$
$- \tau_{0_2}\begin{pmatrix} (S\delta_0)^2 - \boldsymbol{\nu}_0^T\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\nu}_0 & (\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\nu}_0)^T \\ \boldsymbol{\Lambda}_0^{-1}\boldsymbol{\nu}_0 & -\boldsymbol{\Lambda}_0^{-1} \end{pmatrix} \succeq 0, \ \forall \ l = 1, ..., L,$ where $\tau_{0_1}, \tau_{0_2} \geq 0$. Thus, we have that (7.2) is an SCP with probability ambiguity, and an SDP without. □

7.2 Experimental Results

We run some numerical experiments to see what the effect of adding the black swan region to our fragmentation is in terms of negating undesirable effects from unpredictable events. We consider portfolio optimization under conditional value-at-risk as in Section 6.3. We consider the same setup for the portfolio selection problem as in the previous section, which we summarize here. We gather historical data of daily returns for 49 assets during the period starting January 2007 and ending June 2009 available from the Center for Research in Security Prices. We choose this period to include the U.S. financial crisis, as this would be considered a black swan event. We perform 30 experiments (each with different assets, selected randomly) and averag over them. In each experiment, we select four assets and the decision is to select a portfolio from these four assets for each day in the time period. The objective is to make a portfolio allocation $\mathbf{x}$ to minimize the conditional value-at-risk, where $g(\mathbf{x}, \boldsymbol{\xi}) = -\boldsymbol{\xi}^T\mathbf{x}$. We also keep track of the daily returns. On a given day, the decision maker observes a sample of $N = 120$ realizations of the most recent daily returns $\boldsymbol{\xi}_j^i$, for assets $j = 1, ..., p$, and days $i = 1, ..., N$. Given the sample, we compare two versions of RF. For the standard RF approach, we use a fixed $r = 4$, and cluster according to constrained k-means as described in the previous section. Then we solve (6.5) where the Θ is constructed according to the conditional moments estimates. For RF with black swan region added, we construct an additional region Ω_0 as described in the previous section, with $p_0 = .01$ and $S = 10$, and solve (7.2). For simplicity, we do not construct an ambiguity set for

Table 7.1: Portfolio Optimization with Black Swan

	$N = 120$, $p_0 = .01$, S=10	
	RF	**RF with Black Swan**
Average Conditional Value-at-Risk	1.0952	1.0935
Average Daily Return	-0.0049	-0.0046
Average Accumulated Wealth at End of Period	0.9719	0.9732

the probability vector. We continue this procedure for every day over the considered time period.

The results are given in Table 7.1. The out-of-sample CVaR, or average CVaR, is the negative of the average return out of the $1 - \alpha$ worst daily returns in each experiment, where $\alpha = .95$, and then averaged over all experiments. The mean return is computed over all days for each experiment, and then averaged over experiments. The third row contains average accumulated wealth over the time period. That is, starting with a standardized wealth of 1 on day 1, this measures the amount of wealth a user would have, on average (over experiments), at the end of the two-and-a-half year period, if they invested each day in a portfolio following the optimal solution to that day's RF or black swan modified RF strategy.

As can be seen in the table, the numbers for the two columns are very similar. Much of the time, the two solutions are in fact the same. However, over the long term, adding the black swan creates a small benefit. The performance is slightly better in terms of the conditional value-at-risk, average cost, and accumulated return. However, the difference is very slight.

In Figure 7.3, we track the average wealth over each day through the period for a different set of experiments (same setup, different realizations). RF and RF with black swan follow very similar tracks, with the black swan being slightly less prone to extreme drops during catastrophic events. This may indicate that the black swan method is more robust to these types of unpredictable, high impact events. However, over the long term there is little difference in performance. In Figure 7.4, we plot the best and worst experiments out of the 30 for each method. On the extreme ends, the black swan method appears to be superior. Especially in the worst case scenario, the black swan solution is more robust during the financial crisis of 2008.

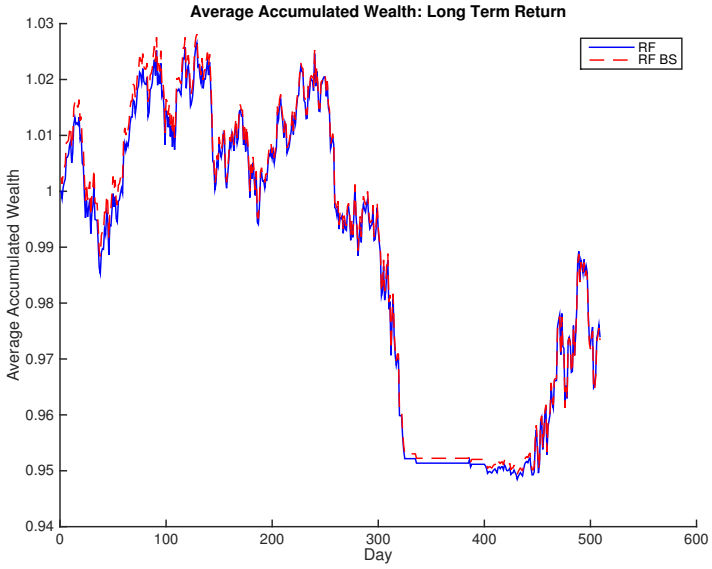

Figure 7.3: Average Accumulated Wealth

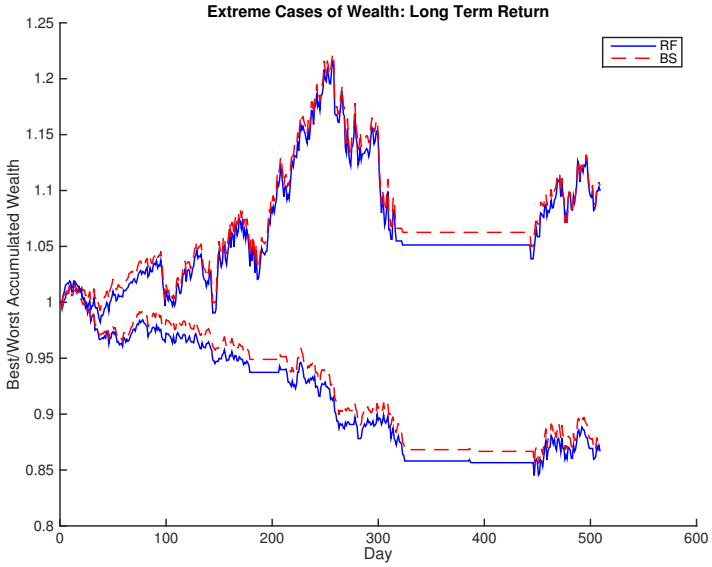

Figure 7.4: Extreme Cases, Accumulated Wealth

Chapter 8

Two-Stage Stochastic Programs

As an extension to our data-driven approach to decision making under distributional ambiguity, we consider problems where the decision must be made in two stages. In this scenario, a decision maker must make an initial decision in the first stage, after which a random event occurs. Following this event, the decision maker responds with a second stage decision. The decision in the first stage must be made without knowledge of the random outcome, which will affect the second stage. The goal is to develop an optimal policy which includes a single first stage decision that takes into account all of the possible second stage scenarios and a collection of second stage decisions, depending upon the outcome of the random event.

We will consider two-stage linear stochastic programming problems. The constraints on the first stage decision vector, $\mathbf{x}$, will be linear. We will denote these linear constraints, $\mathbf{A}\mathbf{x} \leq \mathbf{b}$, as $\mathbf{x} \in \mathscr{X}$ for notational convenience ($\mathscr{X}$ is a polytope). Thus, the two stage linear program can be formulated as:

$$\min_{\mathbf{x}\in\mathscr{X}} \quad \mathbf{c}^T\mathbf{x} + \mathbb{E}[Q(\mathbf{x}, \Lambda)], \tag{8.1}$$

where $Q(\mathbf{x}, \Lambda)$ is the optimal value of the second stage problem:

$$\begin{aligned} Q(\mathbf{x}, \Lambda) = \quad & \min_{\mathbf{w}} \quad \boldsymbol{\xi}(\Lambda)^T\mathbf{w}, \\ & \text{subject to} \quad \mathbf{W}(\Lambda)\mathbf{w} = \mathbf{h}(\Lambda) - \mathbf{T}(\Lambda)\mathbf{x}, \\ & \qquad\qquad\ \ \mathbf{w} \geq 0. \end{aligned} \tag{8.2}$$

Here $\mathbf{x}$ is the first stage decision vector, $\mathbf{w}$ is the second stage decision vector, and random vector Λ contains the information about the data in the second stage, which may include any of $(\boldsymbol{\xi}, \mathbf{W}, \mathbf{h}, \mathbf{T})$. We will consider two cases, one where the randomness occurs in the objective coefficients ($\Lambda = \Xi$) and one where the randomness occurs on the right hand side in the constraints ($\Lambda = H$).

In this formulation, we implicitly assume that we wish to minimize in terms of expectation. That is, to minimize the sum of the first stage cost and the expected second stage cost. As in single stage stochastic programming, this assumes that the distribution of the random vector Λ is known. However, as we argued before, this is often not the case. Even if it is known, it is necessary to assume that Λ has a finite number of possible realizations, or scenarios, in order to solve the problem. Thus, the distribution must be approximated by a discrete distribution. We consider a situation where a data sample is available. One approach would be to use the empirical distribution to approximate, as in SAA, which is referred to as scenario construction. Here the expectation over the second stage is approximated as a sample average, where each term in the sum is a deterministic linear program. Thus, the full two-stage problem can be rewritten as a single, larger, deterministic equivalent where we have a different vector of second stage variables for each scenario.

Alternative solution methods include using the moments to construct an ambiguity set and considering the worst case expectation for the second stage, or employing robust fragmentation in an analogous way as done with single stage stochastic programming. In this section, we will explore tractable formulations for RF on two stage problems.

8.1 Uncertainty in Objective

First we consider the situation in which the objective coefficients of the second stage are random, and all other data is constant. Thus we have:

$$\min_{\mathbf{x}\in\mathscr{X}} \quad \mathbf{c}^T\mathbf{x} + \sup_{\pi\in\Theta} \mathbb{E}_\pi[Q(\mathbf{x}, \Xi)], \tag{8.3}$$

where π lies in the space of probability measures Θ over the measurable space $(\Omega, \mathscr{F})$ and the second stage optimal value is given by:

$$\begin{aligned} Q(\mathbf{x}, \Xi) = \quad & \min_{\mathbf{w}} \quad \Xi^T \mathbf{w}, \\ & \text{subject to} \quad \mathbf{W}\mathbf{w} = \mathbf{h} - \mathbf{T}\mathbf{x}, \\ & \qquad\qquad\quad \mathbf{w} \geq 0. \end{aligned} \tag{8.4}$$

Specifically, we consider an ambiguity set of distributions Θ as before, constrained by regional probabilities and first and second order conditional moments:

$$\begin{aligned} \min_{\mathbf{x} \in \mathscr{X}} \ & \mathbf{c}^T \mathbf{x} + \sup_{\pi \in \Theta} \ \mathbb{E}_\pi [Q(\mathbf{x}, \Xi)], \\ \text{subject to} \ & \mathbb{P}_\pi [\Xi \in \Omega_k] = p_k, \\ & \mathbb{E}_\pi [\Xi | \Xi \in \Omega_k] = \hat{\boldsymbol{\mu}}_k, \\ & \mathbb{E}_\pi [\Xi \Xi^T | \Xi \in \Omega_k] = \hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^T, \quad \text{for } k = 1, ..., K, \\ & \mathbf{e}^T \mathbf{p} = 1, \\ & \mathbf{p} \geq 0, \\ & \sum_{k=1}^{K} \hat{p}_k \phi(\frac{p_k}{\hat{p}_k}) \leq \rho, \end{aligned} \tag{8.5}$$

Theorem 10. *The solution to problem* (8.5)*, when Assumption 2 is satisfied, is given*

by the minimizer of the following optimization problem:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{z_k,\mathbf{z}_k,\mathbf{Z}_k,\mathbf{w}_k\}} \quad & \mathbf{c}^T\mathbf{x}+\eta+\rho\lambda+\lambda\sum_{k=1}^{r}\hat{p}_k\phi^\star\left(\frac{z_k+\mathbf{z}_k{}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta}{\lambda}\right)\\
& +\lambda\sum_{k=r+1}^{K}\hat{p}_k\phi^\star\left(\frac{\mathbf{m}_k^T\mathbf{w}_k-\eta}{\lambda}\right),\\
\textit{subject to} \quad & \boldsymbol{\xi}^T\mathbf{Z}_k\boldsymbol{\xi}+\mathbf{z}_k^T\boldsymbol{\xi}+z_k+(-\mathbf{h}+\mathbf{T}\mathbf{x})^T\mathbf{U}^\dagger\boldsymbol{\xi}\geq 0,\forall\boldsymbol{\xi}\in\Omega_k,\ \forall k=1,...,r,\\
& (-\mathbf{h}+\mathbf{T}\mathbf{x})^T\mathbf{U}^\dagger\leq 0,\\
& z_k+\mathbf{z}_k{}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\forall k=1,...,r,\\
& \mathbf{m}_k^T\mathbf{w}_k-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\ \forall k=r+1,...,K,\\
& \mathbf{W}\mathbf{w}_k=\mathbf{h}-\mathbf{T}\mathbf{x},\ \forall k=r+1,...,K,\\
& \mathbf{w}_k\geq 0,\ \forall k=r+1,...,K,\\
& \lambda\geq 0.
\end{aligned}
\tag{8.6}
$$

Proof. The proof is given in Appendix B.5. □

We have now restructured the infinite dimensional constraints $\boldsymbol{\xi}^T\mathbf{Z}_k\boldsymbol{\xi}+\mathbf{z}_k^T\boldsymbol{\xi}+z_k+(-\mathbf{h}+\mathbf{T}\mathbf{x})^T\mathbf{U}^\dagger\boldsymbol{\xi}\geq 0,\forall\boldsymbol{\xi}\in\Omega_k$ as non-negativity of a function that is quadratic in $\boldsymbol{\xi}$ and linear in $\mathbf{x}$ over each domain Ω_k. All of the other constraints are linear. Thus, all of the same results for the single stage stochastic program apply to the two stage linear program as well. That is, the same techniques can be used for a fragmentation of ellipsoids and hyperellipsoids as in Theorem 3, or for polytopes as in Theorem 2. For

the ellipsoidal case, we get the following formulation:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{\mathbf{w}_k,\tau_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \mathbf{c}^T\mathbf{x}+\eta+\rho\lambda+\lambda\sum_{k=r+1}^{K}\hat{p}_k\phi^\star\left(\frac{\mathbf{m}_k^T\mathbf{w}_k-\eta}{\lambda}\right) \\
& +\lambda\sum_{k=1}^{r}\hat{p}_k\phi^\star\left(\frac{z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta}{\lambda}\right), \\
\text{subject to} \quad & \begin{pmatrix} z_k & \frac{1}{2}(\mathbf{z}_k^T+(-\mathbf{h}+\mathbf{T}\mathbf{x})^T\mathbf{U}^\dagger) \\ \frac{1}{2}(\mathbf{z}_k+\mathbf{W}^\dagger(-\mathbf{h}+\mathbf{T}\mathbf{x})) & \mathbf{Z}_k \end{pmatrix} \\
& -\tau_k\begin{pmatrix} \delta_k^2-\boldsymbol{\nu}_k^T\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & (\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k)^T \\ \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & -\boldsymbol{\Lambda}_k^{-1} \end{pmatrix}\succeq 0,\ \forall k\in E, \\
& \begin{pmatrix} z_k & \frac{1}{2}(\mathbf{z}_k^T+(-\mathbf{h}+\mathbf{T}\mathbf{x})^T\mathbf{U}^\dagger) \\ \frac{1}{2}(\mathbf{z}_k+\mathbf{W}^\dagger(-\mathbf{h}+\mathbf{T}\mathbf{x})) & \mathbf{Z}_k \end{pmatrix} \\
& -\tau_{k_1}\begin{pmatrix} -\delta_{kl}^2+\boldsymbol{\nu}_k^T\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & -(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k)^T \\ -\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & \boldsymbol{\Lambda}_k^{-1} \end{pmatrix} \\
& -\tau_{k_2}\begin{pmatrix} \delta_{ku}^2-\boldsymbol{\nu}_k^T\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & (\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k)^T \\ \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & -\boldsymbol{\Lambda}_k^{-1} \end{pmatrix}\succeq 0, \forall k\in D, \\
& \tau_k\geq 0,\ \forall k\in E, \\
& \tau_{k_1},\tau_{k_2}\geq 0,\ \forall k\in D, \\
& (-\mathbf{h}+\mathbf{T}\mathbf{x})^T\mathbf{U}^\dagger\leq 0, \\
& z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda, \\
& \forall k=1,...,r, \\
& \mathbf{m}_k^T\mathbf{w}_k-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\ \forall k=r+1,...,K, \\
& \mathbf{W}\mathbf{w}_k=\mathbf{h}-\mathbf{T}\mathbf{x},\ \forall k=r+1,...,K, \\
& \mathbf{w}_k\geq 0,\ \forall k=r+1,...,K, \\
& \lambda\geq 0.
\end{aligned}
\tag{8.7}
$$

For a polytopal fragmentation, for each k, we construct an n+1-dimensional polytope Ω_k' from Ω_k by adding another dimension and restricting its value to 1, as before. We let the vertices and rays of Ω_k' be $\mathbf{v}_{ks}$ for $s=1,...,S$ and $\mathbf{r}_{kr}$ for $r=1,...,R$,

respectively. Let $\bar{\mathbf{Z}}_k = \begin{pmatrix} z_k & \frac{1}{2}(\mathbf{z}_k^T + (-\mathbf{h}+\mathbf{T}\mathbf{x})^T\mathbf{U}^\dagger) \\ \frac{1}{2}(\mathbf{z}_k + \mathbf{W}^\dagger(-\mathbf{h}+\mathbf{T}\mathbf{x})) & \mathbf{Z}_k \end{pmatrix}$. The kth infinite dimensional constraint $\begin{pmatrix} 1 & \boldsymbol{\xi}^T \end{pmatrix} \bar{\mathbf{Z}}_k \begin{pmatrix} 1 \\ \boldsymbol{\xi} \end{pmatrix} \geq 0,\ \forall \boldsymbol{\xi} \in \Omega_k$, is equivalent to $\bar{\mathbf{Z}}_k \in \mathcal{CO}(\Omega_k^{'})$. As argued in the proof of Theorem 2, the constraint $\bar{\mathbf{Z}}_k \in \mathcal{CO}(\Omega_k^{'})$ can be replaced with the equivalent condition that $\mathbf{a}^T\bar{\mathbf{Z}}_{kl}\mathbf{a} \geq 0$ for all $\mathbf{a} = \sum_{s=1}^{S} \lambda_s \mathbf{v}_{ks} + \sum_{r=1}^{R} \gamma_r \mathbf{r}_{kr}$, $\lambda_s, \gamma_r \geq 0$. This can be written as

$$\begin{pmatrix} \lambda_1 & \cdots & \lambda_S & \gamma_1 & \cdots & \gamma_R \end{pmatrix} \begin{pmatrix} \mathbf{v}_{k1}^T \\ \vdots \\ \mathbf{v}_{kS}^T \\ \mathbf{r}_{k1}^T \\ \vdots \\ \mathbf{r}_{kR}^T \end{pmatrix} \bar{\mathbf{Z}}_{kl} \begin{pmatrix} \mathbf{v}_{k1} & \cdots & \mathbf{v}_{kS} & \mathbf{r}_{k1} & \cdots & \mathbf{r}_{kR} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_S \\ \gamma_1 \\ \vdots \\ \gamma_R \end{pmatrix} \geq 0$$

for all $\boldsymbol{\lambda}, \boldsymbol{\gamma} \geq 0$ or equivalently that

$$\begin{pmatrix} \mathbf{v}_{k1}^T \\ \vdots \\ \mathbf{v}_{kS}^T \\ \mathbf{r}_{k1}^T \\ \vdots \\ \mathbf{r}_{kR}^T \end{pmatrix} \bar{\mathbf{Z}}_k \begin{pmatrix} \mathbf{v}_{k1} & \cdots & \mathbf{v}_{kS} & \mathbf{r}_{k1} & \cdots & \mathbf{r}_{kR} \end{pmatrix} \in \mathcal{CO}(\mathbb{R}_{U+V}^{+}).$$

Thus, we have:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{\mathbf{w}_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \mathbf{c}^T\mathbf{x}+\eta+\rho\lambda+\lambda\sum_{k=1}^{r}\hat{p}_k\phi^\star\left(\frac{z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta}{\lambda}\right)\\
& +\lambda\sum_{k=r+1}^{K}\hat{p}_k\phi^\star\left(\frac{\mathbf{m}_k^T\mathbf{w}_k-\eta}{\lambda}\right),\\
\text{subject to} \quad & \begin{pmatrix}\mathbf{v}_{k1}^T\\ \vdots\\ \mathbf{v}_{kV}^T\\ \mathbf{r}_{k1}^T\\ \vdots\\ \mathbf{r}_{kU}^T\end{pmatrix}\bar{\mathbf{Z}}_k\begin{pmatrix}\mathbf{v}_{k1} & \cdots & \mathbf{v}_{kV} & \mathbf{r}_{k1} & \cdots & \mathbf{r}_{kU}\end{pmatrix}\in\mathcal{CO}(\mathbb{R}^+_{U+V}),\\
& \forall k=1,...,r,\\
& (-\mathbf{h}+\mathbf{T}\mathbf{x})^T\mathbf{U}^\dagger\leq 0,\\
& z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\\
& \forall k=1,...,r,\\
& \mathbf{m}_k^T\mathbf{w}_k-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\ \forall k=r+1,...,K,\\
& \mathbf{W}\mathbf{w}_k=\mathbf{h}-\mathbf{T}\mathbf{x},\ \forall k=r+1,...,K,\\
& \mathbf{w}_k\geq 0,\ \forall k=r+1,...,K,\\
& \lambda\geq 0.
\end{aligned}
\tag{8.8}
$$

8.2 Production-transportation Problem

An example application of the two stage stochastic program with uncertainty in the objective coefficients of the second stage is the production-transportation problem. Suppose our goal is to minimize total costs while meeting customer orders for a product. There are two types of costs: production costs and transportation costs. We have m

facilities at which to produce and n customer locations. The cost to produce at facility i is given by c_i. The demand at each customer location is d_j. We assume normalized production capabilities of 1 for each facility. For feasibility, we require that $\sum_j d_j < m$. The transportation cost from facility i to customer j is s_{ij}. The decision variables will the production amounts $\mathbf{x}$ and the transportation quantities $\mathbf{W}$. The deterministic version of the problem is as follows:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{W}} \quad & \sum_{i=1}^{m} c_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{n} s_{ij} w_{ij}, \\ \text{subject to} \quad & \sum_{i=1}^{m} w_{ij} = d_j, \ \forall j, \\ & \sum_{j=1}^{n} w_{ij} = x_i, \ \forall i, \\ & 0 \leq x_i \leq 1, \ \forall i, \\ & w_{ij} \geq 0, \ \forall i, j. \end{aligned} \tag{8.9}$$

In the two stage version of the problem, the costs s_{ij} will be random. The first stage decision will be the amount to produce at each facility, x_i. The second stage decision will be the amount to transport, w_{ij}, from each facility to each customer, once the transportation costs are known. Therefore, the two stage RF version will be:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbf{c}^T \mathbf{x} + \sup_{\pi \in \Theta} \mathbb{E}_\pi[Q(\mathbf{x}, S)], \\ \text{subject to} \quad & 0 \leq x_i \leq 1, \ \forall i, \end{aligned} \tag{8.10}$$

where π lies in the space of probability measures Θ over the measurable space $(\Omega, \mathscr{F})$ and the second stage optimal value is given by:

$$\begin{aligned} Q(\mathbf{x}, S) = \quad & \min_{\mathbf{w}} \quad S^T \mathbf{w}, \\ & \text{subject to} \quad \sum_{i=1}^{m} w_{ij} = d_j, \ \forall j, \\ & \qquad\qquad \sum_{j=1}^{n} w_{ij} = x_i, \ \forall i, \\ & \qquad\qquad w_{ij} \geq 0, \ \forall i, j, \end{aligned} \tag{8.11}$$

where both S and $\mathbf{w}$ have been vectorized. Clearly, the randomness occurs in the objective coefficients and this example falls under the category discussed in Section 8.1. The formulations derived for both ellipsoidal and polytopal fragmentations will be valid for this example. As we revert to the dual, we will denote the deterministic version of S as $\mathbf{s}$. Let $\mathbf{W}_1$ be a matrix such that $\mathbf{W}_1\mathbf{w} = \mathbf{d}$. For example, the first row of $\mathbf{W}_1$ will contain a 1 in each position corresponding to $j = 1$ in $\mathbf{w}$ and a 0 everywhere else. Let $\mathbf{W}_2$ be a matrix such that $\mathbf{W}_2\mathbf{w} = \mathbf{x}$. For example, the first row of $\mathbf{W}_2$ will have a 1 in each position corresponding to $i = 1$ in $\mathbf{w}$ and a 0 everywhere else. Then let $\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix}$, so that $\mathbf{W}\mathbf{w} = \begin{pmatrix} \mathbf{d} \\ \mathbf{x} \end{pmatrix}$. We then have constraints of the type $\mathbf{W}\mathbf{w} = \mathbf{h} - \mathbf{T}\mathbf{x}$, where $\mathbf{h} = \begin{pmatrix} \mathbf{d} \\ \mathbf{0} \end{pmatrix}$ and $\mathbf{T} = -\begin{pmatrix} \mathbf{0} \\ \mathbf{I} \end{pmatrix}$. Now we can derive the ellipsoidal fragmentation as in (8.7):

$$\min_{\mathbf{x},\eta,\lambda,\{\mathbf{w}_k,\tau_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad \mathbf{c}^T\mathbf{x}+\eta+\rho\lambda+\lambda\sum_{k=1}^{r}\hat{p}_k\phi^\star\left(\frac{z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\cdot(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta}{\lambda}\right)$$
$$+\lambda\sum_{k=r+1}^{K}\hat{p}_k\phi^\star\left(\frac{\mathbf{m}_k^T\mathbf{w}_k-\eta}{\lambda}\right),$$

subject to

$$\begin{pmatrix} z_k & \frac{1}{2}\left(\mathbf{z}_k^T+\left[-\begin{pmatrix}\mathbf{d}\\\mathbf{0}\end{pmatrix}-\begin{pmatrix}\mathbf{0}\\\mathbf{I}\end{pmatrix}\mathbf{x}\right]^T\mathbf{U}^\dagger\right)\\ \frac{1}{2}\left(\mathbf{z}_k+\mathbf{W}^\dagger\left[-\begin{pmatrix}\mathbf{d}\\\mathbf{0}\end{pmatrix}-\begin{pmatrix}\mathbf{0}\\\mathbf{I}\end{pmatrix}\mathbf{x}\right]\right) & \mathbf{Z}_k\end{pmatrix}$$
$$-\tau_k\begin{pmatrix}\delta_k^2-\boldsymbol{\nu}_k^T\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & (\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k)^T\\ \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & -\boldsymbol{\Lambda}_k^{-1}\end{pmatrix}\succeq 0,\ \forall k\in E,$$

$$\begin{pmatrix} z_k & \frac{1}{2}\left(\mathbf{z}_k^T+\left[-\begin{pmatrix}\mathbf{d}\\\mathbf{0}\end{pmatrix}-\begin{pmatrix}\mathbf{0}\\\mathbf{I}\end{pmatrix}\mathbf{x}\right]^T\mathbf{U}^\dagger\right)\\ \frac{1}{2}\left(\mathbf{z}_k+\mathbf{W}^\dagger\left[-\begin{pmatrix}\mathbf{d}\\\mathbf{0}\end{pmatrix}-\begin{pmatrix}\mathbf{0}\\\mathbf{I}\end{pmatrix}\mathbf{x}\right]\right) & \mathbf{Z}_k\end{pmatrix}$$
$$-\tau_{k_1}\begin{pmatrix}-\delta_{kl}^2+\boldsymbol{\nu}_k^T\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & -(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k)^T\\ -\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & \boldsymbol{\Lambda}_k^{-1}\end{pmatrix}$$
$$-\tau_{k_2}\begin{pmatrix}\delta_{ku}^2-\boldsymbol{\nu}_k^T\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & (\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k)^T\\ \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & -\boldsymbol{\Lambda}_k^{-1}\end{pmatrix}\succeq 0,\forall k\in D,$$

$$\tau_k\geq 0,\ \forall k\in E,\ \tau_{k_1},\tau_{k_2}\geq 0,\ \forall k\in D,\ 0\leq x_i\leq 1,\ \forall i,$$

$$\left[-\begin{pmatrix}\mathbf{d}\\\mathbf{0}\end{pmatrix}-\begin{pmatrix}\mathbf{0}\\\mathbf{I}\end{pmatrix}\mathbf{x}\right]\mathbf{U}^\dagger\leq 0,$$

$$z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\cdot(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\ \forall k=1,...,r,$$

$$\mathbf{m}_k^T\mathbf{w}_k-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\ \forall k=r+1,...,K,$$

$$\mathbf{W}\mathbf{w}_k=\begin{pmatrix}\mathbf{d}\\\mathbf{0}\end{pmatrix}+\begin{pmatrix}\mathbf{0}\\\mathbf{I}\end{pmatrix}\mathbf{x},\ \forall k=r+1,...,K,$$

$$\mathbf{w}_k\geq 0,\ \forall k=r+1,...,K,\ \lambda\geq 0.$$

8.3 Multiple Stages

The formulation derived in Section 8.1 can be extended to the case where there are more than two stages. We again consider the situation in which the objective coefficients of the stages beyond the first are random, and all other data is constant. Thus we have:

$$\min_{\mathbf{x}\in\mathscr{X}} \quad \mathbf{c}^T\mathbf{x} + \sup_{\pi_2\in\Theta_2} \mathbb{E}_{\pi_2}[Q_2(\mathbf{x},\Xi_2)], \tag{8.12}$$

where π_2 lies in the space of probability measures Θ_2 over the measurable space $(\Omega_2, \mathscr{F})$. Let M be the total number of stages. The tth stage is given by:

$$\begin{aligned} Q_t(\mathbf{x},\Xi_t) = \quad & \min_{\mathbf{w}_t} \quad \Xi_t^T\mathbf{w}_t + \sup_{\pi_{t+1}\in\Theta_{t+1}} \mathbb{E}_{\pi_{t+1}}[Q_{t+1}(\mathbf{x},\Xi_{t+1})], \\ & \text{subject to} \quad \mathbf{W}_t\mathbf{w}_t = \mathbf{h}_t - \mathbf{T}_t\mathbf{w}_{t-1}, \\ & \qquad\qquad \mathbf{w}_t \geq 0 \end{aligned} \tag{8.13}$$

for $t = 2, ..., M-1$, where $\mathbf{w}_1 = \mathbf{x}$. The last stage is given by:

$$\begin{aligned} Q_t(\mathbf{x},\Xi_M) = \quad & \min_{\mathbf{w}_M} \quad \Xi_M^T\mathbf{w}_M, \\ & \text{subject to} \quad \mathbf{W}_M\mathbf{w}_M = \mathbf{h}_M - \mathbf{T}_M\mathbf{w}_{M-1}, \\ & \qquad\qquad \mathbf{w}_M \geq 0. \end{aligned} \tag{8.14}$$

We assume the variables Ξ_t in each stage are independent. We have a fragmentation of each support Ω_t, in each stage t, into regions $\Omega_t^1, ..., \Omega_t^{K_t}$. We leave out probability ambiguity, and assume that all regions are non-singletons. Therefore, we are considering the following problem:

$$\begin{aligned} \min_{\mathbf{x}\in\mathscr{X}} \quad & \mathbf{c}^T\mathbf{x} + \sup_{\pi_2\in\Theta_2} \mathbb{E}_{\pi_2}[Q_2(\mathbf{x},\Xi_2)], \\ \text{subject to} \quad & \mathbb{P}_{\pi_2}[\Xi \in \Omega_2^k] = \hat{p}_2^k, \\ & \mathbb{E}_{\pi_2}[\Xi|\Xi \in \Omega_2^k] = \hat{\boldsymbol{\mu}}_2^k, \\ & \mathbb{E}_{\pi_2}[\Xi\Xi^T|\Xi \in \Omega_2^k] = \hat{\boldsymbol{\Sigma}}_2^k + \hat{\boldsymbol{\mu}}_2^k(\hat{\boldsymbol{\mu}}_2^k)^T, \quad \text{for } k = 1, ..., K_2, \end{aligned} \tag{8.15}$$

where $Q_2(\mathbf{x}, \Xi_2)$ is defined as in (8.13), as is every subsequent stage. In terms of notation, we will refer to each kth regional tth stage problem with decision variable $\mathbf{w}_t^k$ as follows:

$$\begin{aligned} Q_t(\mathbf{x}, \boldsymbol{\xi}_t) = \quad & \min_{\mathbf{w}_t^k} \quad \boldsymbol{\xi}_t^T \mathbf{w}_t^k, \\ & \text{subject to} \quad \mathbf{W}_t (\mathbf{w}_t^k)^T = \mathbf{h}_t - \mathbf{T}_t \mathbf{x}, \\ & \qquad\qquad\quad \mathbf{w}_t^k \geq 0. \end{aligned} \tag{8.16}$$

Theorem 11. *The solution to problem* (8.15)*, where each stage after the first is defined as in* (8.13) *and* (8.14)*, is given by the minimizer of the following optimization problem:*

$$\begin{aligned} \min_{\mathbf{x} \in \mathscr{X}, \{z_k, \mathbf{z}_k, \mathbf{Z}_k, \mathbf{w}_k\}} \quad & \mathbf{c}^T \mathbf{x} + \sum_{k=1}^{K_2} \hat{p}_2^k \left(z_2^k + (\mathbf{z}_2^k)^T \hat{\boldsymbol{\mu}}_2^k + \mathbf{Z}_2^k \circ (\hat{\boldsymbol{\Sigma}}_2^k + \hat{\boldsymbol{\mu}}_2^k (\hat{\boldsymbol{\mu}}_2^k)^T) \right), \\ \textit{subject to} \quad & \boldsymbol{\xi}_t^T \mathbf{Z}_t^k \boldsymbol{\xi}_t + (\mathbf{z}_t^k)^T \boldsymbol{\xi}_t + z_t^k + (-\mathbf{h}_t + \mathbf{T}_t \mathbf{x})^T \mathbf{U}_t^\dagger \boldsymbol{\xi}_t \geq \\ & \sum_{k=1}^{K_{t+1}} \hat{p}_{t+1}^k \left[z_{t+1}^k + (\boldsymbol{\mu}_{t+1}^k)^T \mathbf{z}_{t+1}^k + \mathbf{Z}_{t+1}^k \circ (\boldsymbol{\Sigma}_{t+1}^k + \boldsymbol{\mu}_{t+1}^k (\boldsymbol{\mu}_{t+1}^k)^T) \right], \\ & \forall \boldsymbol{\xi}_t \in \Omega_t^k, \ \forall k = 1, ..., K_t, \ \forall t = 2, ..., M-1, \\ & \boldsymbol{\xi}_M^T \mathbf{Z}_M^k \boldsymbol{\xi}_M + (\mathbf{z}_M^k)^T \boldsymbol{\xi}_M + z_M^k + (-\mathbf{h}_M + \mathbf{T}_M \mathbf{x})^T \mathbf{U}_M^\dagger \boldsymbol{\xi}_M \geq 0, \\ & \forall \boldsymbol{\xi}_M \in \Omega_M^k, \ \forall k = 1, ..., K_M, \\ & (-\mathbf{h}_t + \mathbf{T}_t \mathbf{x})^T \mathbf{U}_t^\dagger \leq 0, \forall t = 2, ..., M. \end{aligned} \tag{8.17}$$

Proof. We may follow the same procedure as in the two stage case to get a formulation similar to (8.6). We must take the dual of each inner maximization, one layer at a time. To begin, we follow the exact same procedure for the second stage as if there were no further stages. The only difference is that we have the additional term, $\sup_{\pi_3} \mathbb{E}[Q_3(\mathbf{x}, \boldsymbol{\xi}_3)]$, in the objective function. Since this term is constant in terms of $\boldsymbol{\xi}_2$ and doesn't involve the second stage variables $\mathbf{w}_2^k$, it simply carries through all of the derivations and ends up in the infinite dimensional constraint as follows: $\boldsymbol{\xi}_2^T \mathbf{Z}_2^k \boldsymbol{\xi}_2 + (\mathbf{z}_2^k)^T \boldsymbol{\xi}_2 + z_2^k + (-\mathbf{h}_2 + \mathbf{T}_2 \mathbf{x})^T \mathbf{U}_2^\dagger \boldsymbol{\xi}_2 \geq \sup_{\pi_3} \ \mathbb{E}[Q_3(\mathbf{x}, \boldsymbol{\xi}_3)], \forall \boldsymbol{\xi}_2 \in \Omega_2^k, \ \forall k = 1, ..., K_2$. Thus, it takes the place of 0 on the right hand side. As before, we have the constraint $(-\mathbf{h}_2 + \mathbf{T}_2 \mathbf{x})^T \mathbf{U}_2^\dagger \leq 0$.

We assume that we have conditional moments constraints on the random vector Ξ_3, as we did on Ξ_2, and they are independent. Then we can take the dual of the right hand side to arrive at:

$$\boldsymbol{\xi}_2^T \mathbf{Z}_2^k \boldsymbol{\xi}_2 + (\mathbf{z}_2^k)^T \boldsymbol{\xi}_2 + z_2^k + (-\mathbf{h}_2 + \mathbf{T}_2\mathbf{x})^T \mathbf{U}_2^\dagger \boldsymbol{\xi}_2 \geq$$
$$\inf_{z_3^l, \mathbf{z}_3^l, \mathbf{Z}_3^l} \sum_l \hat{p}_3^l \left[z_3^l + (\hat{\boldsymbol{\mu}}_3^l)^T \mathbf{z}_3^l + \mathbf{Z}_3^l \circ (\hat{\boldsymbol{\Sigma}}_3^l + \hat{\boldsymbol{\mu}}_3^l (\hat{\boldsymbol{\mu}}_3^l)^T) \right], \forall \boldsymbol{\xi}_2 \in \Omega_2^k, \ \forall k = 1, ..., K_2,$$

where $z_3^l, \mathbf{z}_3^l, \mathbf{Z}_3^l$ must satisfy $\boldsymbol{\xi}_3^T \mathbf{Z}_3^l \boldsymbol{\xi}_3 + (\mathbf{z}_3^l)^T \boldsymbol{\xi}_3 + z_3^l \geq Q_3(\mathbf{x}, \boldsymbol{\xi}_3)$ for any $\boldsymbol{\xi}_3 \in \Omega_3^l$, for any l in the second fragmentation. Iteratively we follow the same procedure on the constraints. Since $Q_3(\mathbf{x}, \boldsymbol{\xi}_3) = \min_{\mathbf{w}_3^l} \boldsymbol{\xi}_3^T \mathbf{w}_3^l + \sup_{\pi_4} \ \mathbb{E}[Q_4(\mathbf{x}, \boldsymbol{\xi}_4)]$, the constraint is equivalent to:

$$\boldsymbol{\xi}_3^T \mathbf{Z}_3^l \boldsymbol{\xi}_3 + (\mathbf{z}_3^l)^T \boldsymbol{\xi}_3 + z_3^l + (-\mathbf{h}_3 + \mathbf{T}_3\mathbf{x})^T \mathbf{U}_3^\dagger \boldsymbol{\xi}_3 \geq \sup_{\pi_4} \ \mathbb{E}[Q_4(\mathbf{x}, \boldsymbol{\xi}_4)], \forall \boldsymbol{\xi}_3 \in$$
$$\Omega_3^l, \ \forall l = 1, ..., K_3.$$

where, once again, we have the constraint $(-\mathbf{h}_3 + \mathbf{T}_3\mathbf{x})^T \mathbf{U}_3^\dagger \leq 0$. Taking the dual as before, we get:

$$\boldsymbol{\xi}_3^T \mathbf{Z}_3^l \boldsymbol{\xi}_3 + (\mathbf{z}_3^l)^T \boldsymbol{\xi}_3 + z_3^l + (-\mathbf{h}_3 + \mathbf{T}_3\mathbf{x})^T \mathbf{U}_3^\dagger \boldsymbol{\xi}_3 \geq$$
$$\inf_{z_4^k, \mathbf{z}_4^k, \mathbf{Z}_4^k} \sum_k \hat{p}_4^k \left[z_4^k + (\hat{\boldsymbol{\mu}}_4^k)^T \mathbf{z}_4^k + \mathbf{Z}_4^k \circ (\hat{\boldsymbol{\Sigma}}_4^k + \hat{\boldsymbol{\mu}}_4^k (\hat{\boldsymbol{\mu}}_4^k)^T) \right], \forall \boldsymbol{\xi}_3 \in \Omega_3^l, \ \forall l = 1, ..., K_3,$$

Thus, we follow this procedure iteratively to arrive at the following constraints:

$$\begin{aligned}
&\boldsymbol{\xi}_t^T \mathbf{Z}_t^k \boldsymbol{\xi}_t + (\mathbf{z}_t^k)^T \boldsymbol{\xi}_t + z_t^k + (-\mathbf{h}_t + \mathbf{T}_t\mathbf{x})^T \mathbf{U}_t^\dagger \boldsymbol{\xi}_t \geq \\
&\sum_{k=1}^{K_{t+1}} \hat{p}_{t+1}^k \left[z_{t+1}^k + (\boldsymbol{\mu}_{t+1}^k)^T \mathbf{z}_{t+1}^k + \mathbf{Z}_{t+1}^k \circ (\boldsymbol{\Sigma}_{t+1}^k + \boldsymbol{\mu}_{t+1}^k (\boldsymbol{\mu}_{t+1}^k)^T) \right], \ \forall \boldsymbol{\xi}_t \in \Omega_t^k, \\
&\forall k = 1, ..., K_t, \ \forall t = 2, ..., M-1, \\
&\boldsymbol{\xi}_M^T \mathbf{Z}_M^k \boldsymbol{\xi}_M + (\mathbf{z}_M^k)^T \boldsymbol{\xi}_M + z_M^k + (-\mathbf{h}_M + \mathbf{T}_M\mathbf{x})^T \mathbf{U}_M^\dagger \boldsymbol{\xi}_M \geq 0, \ \forall \boldsymbol{\xi}_M \in \Omega_M^k, \\
&\forall k = 1, ..., K_M, \\
&(-\mathbf{h}_t + \mathbf{T}_t\mathbf{x})^T \mathbf{U}_t^\dagger \leq 0, \forall t = 2, ..., M.
\end{aligned} \tag{8.18}$$

We can drop the infimums because the constraint of being greater than or equal to the infimum is equivalent to a feasibility constraint; that is, that a set of such parameters exist. If we construct a fragmentation of ellipsoids and differences of ellipsoids, then each of these constraints can be reformulated using S-Lemma or extended S-Lemma, as they are all quadratic in $\boldsymbol{\xi}_t$. Thus, we can formulate the problem as in (8.7), but with

constraints for every $t = 2, ..., M$, and for each $k = 1, ..., K_t$, and with the additional summation term as a constant in each constraint as it does not involve the variable $\boldsymbol{\xi}_t$, which results in (8.17). □

8.4 Uncertainty in Constraints

Now we consider a two-stage stochastic program where the uncertainty is in the vector H in the right-hand side of the second stage, as follows:

$$\begin{aligned} Q(\mathbf{x}, H) = \quad & \min_{\mathbf{w}} \quad \boldsymbol{\xi}^T \mathbf{w}, \\ & \text{subject to} \quad \mathbf{W}\mathbf{w} = H - \mathbf{T}\mathbf{x}, \\ & \qquad\qquad\ \ \mathbf{w} \geq 0. \end{aligned} \tag{8.19}$$

Theorem 12. *The solution to problem* (8.5)*, when the uncertainty is in the constraint vector* $\mathbf{h}$ *as in* (8.19)*, rather than the objective coefficients, and when Assumption 2 is satisfied, and all of the fragmentation regions* Ω_k *are ellipsoids or differences of ellipsoids as defined in Theorem 3, is given by the minimizer of the following optimization problem:*

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{\mathbf{w}_k,\tau_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \eta+\rho\lambda+\lambda\sum_{k=1}^{r}\hat{p}_k\phi^\star\left(\frac{z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta}{\lambda}\right)\\
& +\lambda\sum_{k=r+1}^{K}\hat{p}_k\phi^\star\left(\frac{\boldsymbol{\xi}^T\mathbf{w}_k-\eta}{\lambda}\right),\\
\textit{subject to} \quad & \begin{pmatrix} z_k+\mathbf{x}^T\mathbf{T}^T\mathbf{U}^\dagger\boldsymbol{\xi} & \frac{1}{2}(\mathbf{z}_k^T-\boldsymbol{\xi}^T\mathbf{W}^\dagger)\\ \frac{1}{2}(\mathbf{z}_k-\mathbf{U}^\dagger\boldsymbol{\xi}) & \mathbf{Z}_k\end{pmatrix}\\
& -\tau_k\begin{pmatrix}\delta_k^2-\boldsymbol{\nu}_k^T\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & (\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k)^T\\ \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & -\boldsymbol{\Lambda}_k^{-1}\end{pmatrix}\succeq 0,\ \forall k\in E,\\
& \begin{pmatrix} z_k+\mathbf{x}^T\mathbf{T}^T\mathbf{U}^\dagger\boldsymbol{\xi} & \frac{1}{2}(\mathbf{z}_k^T-\boldsymbol{\xi}^T\mathbf{W}^\dagger)\\ \frac{1}{2}(\mathbf{z}_k-\mathbf{U}^\dagger\boldsymbol{\xi}) & \mathbf{Z}_k\end{pmatrix}\\
& -\tau_{k_1}\begin{pmatrix}-\delta_{kl}^2+\boldsymbol{\nu}_k^T\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & -(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k)^T\\ -\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & \boldsymbol{\Lambda}_k^{-1}\end{pmatrix}\\
& -\tau_{k_2}\begin{pmatrix}\delta_{ku}^2-\boldsymbol{\nu}_k^T\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & (\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k)^T\\ \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k & -\boldsymbol{\Lambda}_k^{-1}\end{pmatrix}\succeq 0,\forall k\in D,\\
& \tau_k\geq 0,\ \forall k\in E,\ \tau_{k_1},\tau_{k_2}\geq 0,\ \forall k\in D,\\
& -\delta_k\sqrt{\boldsymbol{\omega}_i^T\boldsymbol{\Lambda}_k\boldsymbol{\omega}_i}+\boldsymbol{\omega}_i^T\boldsymbol{\nu}_k-\boldsymbol{\omega}_i^T\mathbf{T}\mathbf{x}\geq 0,\forall i=1,...,p,\ \forall k=1,...,r,\\
& z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\\
& \forall k=1,...,r,\\
& \boldsymbol{\xi}^T\mathbf{w}_k-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\ \forall k=r+1,...,K,\\
& \mathbf{W}\mathbf{w}_k=\mathbf{h}_k-\mathbf{T}\mathbf{x},\ \forall k=r+1,...,K,\\
& \mathbf{w}_k\geq 0,\ \forall k=r+1,...,K,\\
& \lambda\geq 0.
\end{aligned}
\tag{8.20}
$$

Proof. The proof is given in Appendix B.6. □

Let $\bar{\mathbf{Z}}_{k0} = \begin{pmatrix} z_k + \mathbf{x}^T\mathbf{T}^T\mathbf{U}^\dagger\boldsymbol{\xi} & \frac{1}{2}(\mathbf{z}_k^T - \boldsymbol{\xi}^T\mathbf{W}^\dagger) \\ \frac{1}{2}(\mathbf{z}_k - \mathbf{U}^\dagger\boldsymbol{\xi}) & \mathbf{Z}_k \end{pmatrix}$, and $\bar{\mathbf{Z}}_{ki} = \begin{pmatrix} -\boldsymbol{\omega}_i^T\mathbf{T}\mathbf{x} & \frac{1}{2}\boldsymbol{\omega}_i^T \\ \frac{1}{2}\boldsymbol{\omega}_i & 0 \end{pmatrix}$ for $i = 1, ..., p$. Then the polytopal fragmentation formulation is as follows:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{\mathbf{w}_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \eta + \rho\lambda + \lambda\sum_{k=1}^{r}\hat{p}_k\phi^\star\left(\frac{z_k + \mathbf{z}_k{}^T\hat{\boldsymbol{\mu}}_k + \mathbf{Z}_k \circ (\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T) - \eta}{\lambda}\right) \\
& + \lambda\sum_{k=r+1}^{K}\hat{p}_k\phi^\star\left(\frac{\boldsymbol{\xi}^T\mathbf{w}_k - \eta}{\lambda}\right), \\
\text{subject to} \quad & \begin{pmatrix} \mathbf{v}_{k1}^T \\ \vdots \\ \mathbf{v}_{kV}^T \\ \mathbf{r}_{k1}^T \\ \vdots \\ \mathbf{r}_{kU}^T \end{pmatrix} \bar{\mathbf{Z}}_{ki} \begin{pmatrix} \mathbf{v}_{k1} & \cdots & \mathbf{v}_{kV} & \mathbf{r}_{k1} & \cdots & \mathbf{r}_{kU} \end{pmatrix} \in \mathcal{CO}(\mathbb{R}^+_{U+V}), \\
& \forall i = 0, ..., p,\ \forall k = 1, ..., r, \\
& z_k + \mathbf{z}_k{}^T\hat{\boldsymbol{\mu}}_k + \mathbf{Z}_k \circ (\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T) - \eta \leq \left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda, \\
& \forall k = 1, ..., r, \\
& \boldsymbol{\xi}^T\mathbf{w}_k - \eta \leq \left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\ \forall k = r+1, ..., K, \\
& \mathbf{W}\mathbf{w}_k = \mathbf{h}_k - \mathbf{T}\mathbf{x},\ \forall k = r+1, ..., K, \\
& \mathbf{w}_k \geq 0,\ \forall k = r+1, ..., K, \\
& \lambda \geq 0.
\end{aligned}
\tag{8.21}
$$

Chapter 9

Conclusion

In this thesis, we present an approach to decision-making in the face of uncertainty, when no distribution is known for the random quantities and only some past data is available. We introduce robust fragmentation, a data-driven approach to approximate a stochastic program using an historical sample. RF is a generalization of previous approaches that allows for the construction of an ambiguity set that is statistically meaningful, consistent with historical data, and results in tractable formulations for some fragmentation types. RF is an extension of the classical MDRO framework that reduces conservativeness and leverages data more effectively. It is also a generalization of SAA that avoids overfitting and reduces problem size. We allow for the introduction of ambiguity in both support and probability. Using RF, we may avoid overfitting without an overly conservative solution. We demonstrate how RF serves as a bridge between MDRO, SAA, and PRO. Balancing the degree of inconsistency of RF with the variability in the solution is analogous to the bias variance tradeoff in statistics and is a core justification for our approach. The ambiguity set may be tailored to the amount, size, and structure of the data.

In RF, the data is broken up into pieces and summarized regionally, reducing the problem size. In essence, the data is reduced in order to construct the ambiguity set, which contains all distributions that would amount to the same reduction. RF performs best for intermediate sample size by dissecting modal information that MDRO fails to account for, but condensing the data to avoid overfitting to improve it upon SAA. For large N, it competes well with SAA and scales much better because of the reduction in

problem size.

RF leads to tractable formulations for many different fragmentations, including univariate, ellipsoidal, and differences of ellipsoids. An extension to the S-Lemma allows us to take advantage of a fragmentation consisting of differences of ellipsoids. We are also able to derive a polynomial size problem for the polytope case. Convergence of the method, when using sample estimates of the moments, to the true solution can be proven. In terms of practical implementation, we outline our algorithm, develop heuristics for determining the fragmentation, and explore clustering techniques. We consider applications for which our method is useful and perform some experiments to demonstrate that it may be a desirable approach in some instances. RF has many applications in inventory control, portfolio optimization under conditional value-at-risk, and facility location, as examples. A new formulation is derived for a facility location problem with L_2 distance. RF performs especially well when the underlying distribution is multimodal or when it is changing over time. We also find that using black swan regions, an adaptation inspired by the concept of an unpredictable but influential event, can help in making decisions for portfolio allocation. We justify RF both under risk neutral decision making via good performance in terms of expected loss and risk aversion through measures of variance in outcome.

We extend the framework to include two-stage stochastic programs. Further investigation into the topic may uncover other forms of stochastic programs for which RF is applicable. Future implementations of the method may introduce an extra element of randomization such as is done in the concept of random forests for decision trees. We could construct a set of fragmentations and determine the robust optimal values for our decision variables and then average over the fragmentations. Fragmentation types, such as polytopes, will become more tractable following any advances in the field of copositive programming. It is desirable to extend the concept to learning problems as well, as RF is a naturally adaptive procedure.

Chapter 10

Bibliography

References

Alpaydin, Ethem. *Introduction to Machine Learning 3rd ed.* The MIT Press. Cambridge, MA 2014.

Ben-Tal, A., A. Nemirovski. (1998) Robust convex optimization. *Math. Oper. Res.* 23:769-805.

Ben-Tal, A., L. El Ghaoui, A. Nemirovski, eds. (2002) Robust Optimization. *Mathematical Programming* Ser. B (92): 453-480.

Ben-Tal, Aharon, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg and Gijs Rennen. (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59.2: 341-357.

D. P. Bertsekas. *Linear Network Optimization.* MIT Press, Cambridge, MA, 1991.

Bertsimas, Dimitris, David B. Brown, and Constantine Caramanis. (2011) Theory and applications of robust optimization. *SIAM Review* 53.3: 464-501.

Bertsimas, Dimitris, Vishal Gupta, and Nathan Kallus. (2013) Data-driven robust optimization. Submitted to *Operations Research*: arXiv 1401.0212.

Bertsimas, Dimitris, Vishal Gupta, and Nathan Kallus. Robust sample average approximation. arXiv:1408.4445 or http://nathankallus.com/

Bertsimas, Dimitris and Ioana Popescu. (2002) On the Relation Between Option and Stock Prices: A Convex Optimization Approach. *Operations Research* 50(2): 358-374.

Bertsimas, Dimitris, Xuan Vinh Doan, Karthik Natarajan, and Chung-Piaw Teo. (2010) Models for Minimax Stochastic Linear Optimization Problems with Risk Aversion. *Mathematics of Operations Research* 35(3): 580-602.

Bertsimas, Dimitris and Ioana Popescu. (2005) Optimal Inequalities in Probability Theory: A Convex Optimization Approach. *SIAM Journal on Optimization* 15(3): 780-804.

Bertsimas, Dimitris and Aurelie Thiele. (2006) A Robust Optimization Approach to Inventory Theory. *Operations Research* 54(1): 150-168.

Birbil, S. Ilker, J.B. Frenk, Joaquim Gromicho, and Shuzhong Zhang. (2009) The role of robust optimization in single-leg airline revenue management. *Management Science* 55.1: 148-163.

Bishop, Christopher M. *Pattern Recognition and Machine Learning*. Springer. 2006.

Borgelt, Christian, and Rudolf Kruse. (2005) Fuzzy and probabilistic clustering with shape and size constraints. Proc. 15th Int. Fuzzy Systems Association World Congress (IFSA05, Beijing, China).

Boyd, Stephen and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press: New York, NY 2004.

Burer, Samuel. (2012) Copositive Programming. Handbook on Semidefinite, Conic, and Polynomial Optimization (Ch. 8). *International Series in Operations Research and Management Science* 166: 201-218.

Burer, Samuel and Hongbo Dong. (2012) Representing Quadratically Constrained Quadratic Programs as Generalized Copositive Programs. *Operations Research Letters* 40: 203-206.

Burer, Samuel. (2009) On the Copositive Representation of Binary and Continuous Nonconvex Quadratic Programs. *Mathematical Programming* 120: 479-495.

Bradley, P.S., K. P. Bennett and A. Demiriz. (2000) Constrained K-Means Clustering. *Technical Report*, Microsoft Research, Redmond: 1-8.

Chen, Li, Simai He and Shuzhong Zhang. (2010) Tight Bounds for Some Risk Measures, with Applications to Robust Portfolio Selection. *Operations Research* 59(4): 847-865.

Daszykowski, Michal. ¡http://chemometria.us.edu.pl/index.php?goto=downloads¿ (Nov 29 2015)

Daszykowski, M., B. Walczak, and D. L. Massart. (2001) Looking for Natural Patterns in Data. Part 1: Density Based Approach. Chemom. Intell. Lab. Syst. 56: 83-92.

Davies, David L. and Donald W. Bouldin. (1979) A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. PAMI-1 (2): 224-227.

Delage, Erick and Yinyu Ye. (2010) Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. *Operations Research* 58(3): 595-612.

Delage, Erick. (2009) Distributionally Robust Optimization in Context of Data-Driven Problems. Dissertation, Stanford University.

DeMiguel, Victor, Lorenzo Garlappi, Francisco J. Nogales and Raman Uppal. (2009) A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science* 55.5: 798-812.

Dempster, A. P., Laird N. M., and Rubin D. B. (1977) Maximum Likelihood from Incomplete

Data via the EM Algorithm. *Journal of the Royal Statistical Society.* Series B (Methodological) 39.1: 1-38.

Diananda, P. H. (1962) On Non-negative Forms in Real Variables Some or All of Which are Non-negative. *Mathematical Proceedings of the Cambridge Philosophical Society* 58(1): 17-25.

Dupacov, J. (1987) The minimax approach to stochastic programming and an illustrative approach. *Stochastics* 20(1): 73-88, 123.

Dupacov, J. (1966) On minimax solutions of stochastic linear programming problems. *Casopis pro p estovn matematiky* 91(4): 423-430.

Dur, Mirjam. Copositive Programming. *Recent Advances in Optimization and its Applications in Engineering*: (ch. 1): 3-20. Springer Publishing: New York, NY 2010.

Epstein, Larry G. (1999) A definition of uncertainty aversion. *The Review of Economic Studies* 66.3: 579-608.

Ester, Martin, Hans-Peter Kriegel, Jiirg Sander, and Xiaowei Xu. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd.* Vol. 96. No. 34.

Gallego, Guillermo and Ilkyeong Moon. (1993) The Distribution Free Newsboy Problem: Review and Extensions. *Journal of the Operational Research Society* 44: 825-834.

Gallego, Guillermo and Ilkyeong Moon. (1994) Distribution Free Procedures for Some Inventory Models. *Journal of the Operational Research Society* 6(1): 651-658.

Gilboa, Itzhak, and David Schmeidler. (1989) Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics* 18.2: 141-153.

Godfrey, Gregory and Warren Powell. (2001) An Adaptive, Distribution-Free Algorithm for the Newsvendor Problem with Censored Demands, with Applications to Inventory and Distribution. *Management Science* 47(8): 1101-1112.

Goh, J., and Sim, M. (2010) Distributionally robust optimization and its tractable approximations. *Operations Research* 58: 902-917.

Grotschel, M., L. Lovasz, A. Schrijver. (1981) The Ellipsoid Method and its Consequences in Combinatorial Optimization. *Combinatorica* 1: 169-197.

Hanasusanto, Grani, Daniel Kuhn, Stein Wallace and Steve Zymler. (2014) Distributionally Robust Multi-Item Newsvendor Problems with Multimodal Demand Distributions. *Mathematical Programming* 152(1): 1-32

Hu, Zhaolin, and L. Jeff Hong. (2013) Kullback-Leibler divergence constrained distributionally robust optimization. *Optimization Online*

http://www.optimization-online.org/DB_FILE/2012/3677.pdf

Isii, K. (1960) The Extrema of Probability Determined by Generalized Moments Bounded Random Variables. *Ann. Inst. Stat. Math.* 12: 119-133.

Krokhmal, Pavlo, Jonas Palmquist, and Stanislav Uryasev. (2002) Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of Risk* 4: 43-68.

Levi, Retsef, Robin Roundy and David Shmoys. (2007) Provably Near-Optimal Sampling- Based Policies for Stochastic Inventory Control Models. *Mathematics of Operations Research* 32(4): 821-839.

Li, Xiaobo, Karthik Natarajan, Chung-Piaw Teo and Zhichao Zheng. (2014) Invited Review: Distributionally Robust Mixed Integer Linear Programs: Persistency Models with Applications. *European Journal of Operational Research* 233(3): 459-473.

Lim, Andrew EB, J. George Shanthikumar, and Gah-Yi Vahn. (2011) Conditional value-at-risk in portfolio optimization: Coherent but fragile. *Operations Research Letters* 39.3: 163-171.

Lloyd, Stuart. (1982) Least squares quantization in PCM. *IEEE transactions on information theory* 28.2s: 129-137.

Love, David, and Guzin Bayraksan. (2013) Two-stage likelihood robust linear program with application to water allocation under uncertainty. Simulation Conference (WSC), Winter. *IEEE*.

Luo, Zhi-Quan, Jos F. Sturm, and Shuzhong Zhang. (2004) Multivariate Nonnegative Quadratic Mappings. *SIAM Journal on Optimization* 14(4): 1140-1162.

Maxfield, John E., and Henryk Minc. (1962) On the matrix equation $X^T X = A$. *Proceedings of the Edinburgh Mathematical Society* (Series 2) 13.02: 125-129.

Mevissen, Martin, Emanuele Ragnoli, and Jia Yuan Yu. (2013) Data-driven Distributionally Robust Polynomial Optimization. *Advances in Neural Information Processing Systems* 26 (NIPS).

Murty, Katta, and Santosh Kabadi. (1987) Some NP-Complete Problems in Quadratic and Nonlinear Programming. *Mathematical Programming* 39: 117-129.

Natarajan, Karthik, Melvyn Sim and Joline Uichanco. (2014) A Data-Driven Heuristic for Inventory Models with Incomplete Demand Information using Robust Partitioning. Submitted to *Management Science*.

Natarajan, Karthik, Dongjian Shi and Kim-Chuan Toh. (2013) Probabilistic Model for Minmax Regret in Combinatorial Optimization. *Operations Research* 62(1): 160-181.

Natarajan, Karthik, Chung-Piaw Teo, and Zhichao Zheng. (2011) Mixed 0-1 Linear Programs

Under Objective Uncertainty: A Completely Positive Representation. *Operations Research* 59(3): 713-728.

Pardo L. *Statistical Inference Based on Divergence Measures.* Chapman and Hall/CRC: Boca Raton, FL. 2006.

Perakis, Georgia and Guillaume Roels. (2008) Regret in the Newsvendor Model with Partial Information. *Operations Research* 56(1): 188-203.

Perakis, Georgia and Guillaume Roels. (2007) The Price of Information: Inventory Management with Limited Information About Demand. Manufacturing and Service. *Operations Management* 8(1): 102-104.

Petrovic, Slobodan. (1996) A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. Proceedings of the 11th Nordic Workshop of Secure IT Systems. 2006. *Kdd* 96(34).

Plik, Imre, and Tams Terlaky. (2007) A survey of the S-lemma. *SIAM Review* 49.3: 371-418.

Popescu, Ioana. (2005) A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Mathematics of Operations Research* 30.3: 632-657.

Popescu, Ioana. (2007) Robust Mean-Covariance Solutions for Stochastic Optimization. *Operations Research* 55(1): 98-112.

Rockafellar, R. Tyrrell, and Stanislav Uryasev. (2000) Optimization of conditional value-at-risk. *Journal of Risk* 2: 21-42.

Scarf, H. A min-max solution of an inventory problem, In: K. J. Arrow, S. Karlin, and H. Scarf, (Eds.), *Studies in the Mathematical Theory of Inventory and Production* Stanford University Press, (1958) 201-209.

Shapiro, Alexander. (2001) On duality theory of conic linear problems. *Semi-infinite Programming.* Nonconvex Optimization and its Applications series (7): 135-165.

Shapiro, A., Ahmed, S. (2004) On a class of minimax stochastic programs. *SIAM J. Optim.* 14(1): 1237-1249.

Shapiro, Alexander, and Darinka Dentcheva. *Lectures on stochastic programming: modeling and theory.* Vol. 16. SIAM and Mathematical Programming Society, 2014.

Shapiro, Alexander, and Tito Homem-de-Mello. (2000) On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs. *SIAM Journal on Optimization* 11.1: 70-86.

Sturm, Jos F. and Shuzhong Zhang. (2003) On Cones of Nonnegative Quadratic Functions. *Mathematics of Operations Research* 28(2): 246-267.

Sun, Hailin, and Huifu Xu. (2015) Convergence analysis for distributionally robust optimization and equilibrium problems. *Mathematics of Operations Research* 41(2): 377-401.

Sun, Peng, and Robert M. Freund. (2004) Computation of minimum-volume covering ellipsoids. *Operations Research* 52.5: 690-706.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie. (2001) Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2: 411-423.

Taleb, Nassim Nicholas. *The Black Swan*. Random House, Incorporated: New York, NY 2007.

VanAntwerp, Jeremy, and Richard Braatz. (2000) A Tutorial of Linear and Bilinear Matrix Inequalities. *Journal of Process Control* 10: 363-385.

Wang, Zizhuo, Peter W. Glynn, and Yinyu Ye. (2013) Likelihood robust optimization for data-driven problems. *Computational Management Science* 13(2): 241-261.

Weisemann, Wolfram, Daniel Kuhn, and Melvyn Sim. (2014) Distributionally Robust Convex Optimization. *Operations Research* 62(6): 1358-1376.

Wong, Man Hong and Shuzhong Zhang. (2014) On Distributional Robust Probability Functions and their Computations. *European Journal of Operational Research* 233: 23-33.

Wu, Chenchen, Donglei Du, and Dachuan Xu. (2015) An Approximation Algorithm for the Two-Stage Distributionally Robust Facility Location Problem. *Advances in Global Optimization*. Springer Proceedings in Mathematics and Statistics (95): 99-107.

Xia, Yong, Shu Wang, and Ruey-Lin Sheu. (2015) S-lemma with equality and its applications. *Mathematical Programming* 156(1): 513-547.

Xu, Huan, Constantine Caramanis and Shie Mannor. (2012) A Distributional Interpretation of Robust Optimization. *Mathematics of Operations Research* 37(1): 95-110.

Yue, Jinfeng, Bintong Chen, and Min-Chiang Wang. (2006) Expected Value of Distribution Information for the Newsvendor Problem. *Operations Research* 54(6): 1128-1136.

Zaiane, Osmar. *Ch. 8: Data Clustering*. Lecture notes, University of Alberta. ¡https://webdocs.cs.ualberta.ca/ zaiane/courses/cmput690/slides/Chapter8/ index.html¿.

Zhu, Zhisu, Jiawei Zhang and Yinyu Ye. (2006) Newsvendor Optimization with Limited Distribution Information. *Optimization Methods and Software* 28(3): 640-667.

Appendix A

Preparatory Material

A.1 Preparatory Material

Lemma 3. *Often used in the field of quadratic optimization, the S-Lemma provides a powerful equivalent condition for the non negativity of a quadratic function $f(\mathbf{x})$ over a quadratic inequality $q(\mathbf{x}) \geq 0$. The theorem states the equivalence of the following conditions:*

$$q(\mathbf{x}) \geq 0 \implies f(\mathbf{x}) \geq 0 \ \ \forall \mathbf{x} \in \mathbb{R}^n \iff \exists\, \tau \geq 0 \ s.t. \ f(\mathbf{x}) - \tau q(\mathbf{x}) \geq 0 \ \ \forall \mathbf{x} \in \mathbb{R}^n, \quad \text{(A.1)}$$

as long as $q(\mathbf{x}_0) > 0$ for some $\mathbf{x}_0$.

Lemma 4. $\mathcal{CO}(\Omega) = \mathcal{CO}(conic(\Omega))$.

Proof. Clearly, $\mathcal{CO}(\Omega) \subseteq \mathcal{CO}(\text{conic}(\Omega))$. For all $\lambda \geq 0$ and $\mathbf{x} \in \Omega$, we have that $(\lambda\mathbf{x})^T A(\lambda\mathbf{x}) = \lambda^2 \mathbf{x}^T \mathbf{A}\mathbf{x} \geq 0$ so that $\mathcal{CO}(\Omega) \supseteq \mathcal{CO}(\text{conic}(\Omega))$. □

Lemma 5. *For $n \leq 4$, $\mathcal{CO}_n = \mathcal{S}_n + \mathcal{N}_n$, where $\mathbf{N}_{(ij)} \geq 0 \ \ \forall \ \ i, j \ \forall \mathbf{N} \in \mathcal{N}_n$. (Maxfield et al. (1962))*

Lemma 6. *The S-Lemma with equality constraints states the equivalence of the following conditions:*

$$q(\mathbf{x}) = 0 \implies f(\mathbf{x}) \geq 0 \ \ \forall \mathbf{x} \in \mathbb{R}^n \iff \exists\, \tau \ s.t. \ f(\mathbf{x}) + \tau q(\mathbf{x}) \geq 0 \ \ \forall \mathbf{x} \in \mathbb{R}^n \quad \text{(A.2)}$$

where f, q are quadratic functions, q is strictly convex, and there exists $\mathbf{x}_0$, $\mathbf{x}_1$ such that $q(\mathbf{x}_0) > 0$ and $q(\mathbf{x}_1) < 0$. (Xia et al. (2015))

A.2 Example Details

Example 1

We consider a one-dimensional newsvendor problem with equal holding and backlog costs, where a manager must determine an inventory level x to minimize a cost $|x - \Xi|$ for the resulting overstock or understock, where Ξ is the unknown demand. Given that Ξ follows some distribution π, the minimizer of the expected cost is the median, m. Given a historical sample of size N, the minimizer of the SAA will be the sample median, $\hat{m}_N$. The solution to a MDRO formulation with sample mean $\hat{\mu}_N$ and sample variance $\hat{\sigma}_N^2$ will be the sample mean, $\hat{\mu}_N$ (as shown in Scarf (1958)). We can easily compute the solution for a PRO version under, for example, variation distance, which we denote as $\hat{\psi}_N$. If we break up the positive real line into two unbounded intervals ($K = 2$) and compute conditional moments, we can solve for the closed-form solution for the RF version, for which we give details in Theorem (7). We denote this solution by $\hat{\nu}_N$. We note that this solution is not even the most optimal RF, but only a very simple one for which we can get a closed form solution.

We consider a distribution π for Ξ that takes values in $\{0, 1, 13, 14, 16\}$, each with equal probability $\frac{1}{5}$. Since π is discrete, so are the distributions for each estimator, with computable pdfs for any given N. Thus, for each of the four solutions $\hat{x}_N$, we can calculate the expected cost over a sample of size N, $\mathbb{E}[|\hat{x}_N - \Xi|]$, where the expectation is over both the random estimator $\hat{x}_N$ and the random demand Ξ. As the objective is to minimize the expected cost $\mathbb{E}[|x - \Xi|]$, we can directly compare the quality of the solutions. The true minimizer is the median, 13, which achieves $\mathbb{E}[|m - \Xi|] = 5.8$. For different values of N, we compute the expected differences $\mathbb{E}[|\hat{\mu}_N - \Xi|]$, $\mathbb{E}[|\hat{m}_N - \Xi|]$, $\mathbb{E}[|\hat{\psi}_N - \Xi|]$, and $\mathbb{E}[|\hat{\nu}_N - \Xi|]$. We compute these values for selected N and plot them in Figure 2.

Note that these are not simulated estimates for the expected cost, but the exact values (except for $N > 10$ where the computational overhead of exactly computing the expected costs becomes prohibitive). We chose the interval boundary to be 12.

Appendix B

Additional Proofs

B.1 Proof of Theorem 1

Proof. The summation term does not involve the inner infinite dimensional variables, π_k. Each inner moment problem is independent as the regions are disjoint and hence we can compute each dual problem independently. We compute these dual problems (see Shapiro (2001) for details about duality theory for conic moment problems) and the dual of the outer robust problem as according to Ben-Tal et al. (2013). Strong duality holds, as shown by Isii (1960), as long as $\hat{\mathbf{\Sigma}}_k \succ 0$ for all k. This condition holds with probability 1 when each $\hat{\mathbf{\Sigma}}_k$ is a regional sample covariance matrix from a continuous generating distribution.

We compute the dual for each kth inner worst case conditional expectation. The relevant constraints are the conditional moments for region k:

$$\begin{aligned} \sup_{\pi_k \in \Theta_k} \quad & \int_{\Omega_k} f(\mathbf{x}, \boldsymbol{\xi}) d\pi_k(\boldsymbol{\xi}), \\ \text{subject to} \quad & \int_{\Omega_k} d\pi_k(\boldsymbol{\xi}) = 1, \\ & \int_{\Omega_k} \boldsymbol{\xi} d\pi_k(\boldsymbol{\xi}) = \hat{\boldsymbol{\mu}}_k, \\ & \int_{\Omega_k} \boldsymbol{\xi}\boldsymbol{\xi}^T d\pi_k(\boldsymbol{\xi}) = \hat{\mathbf{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^T, \ \forall k = 1, ..., r. \end{aligned} \tag{B.1}$$

We associate dual variables z_k, $\mathbf{z}_k$, and $\mathbf{Z}_k$ for the zeroth, first, and second order moments

constraints, respectively. The unconstrained version of the problem is:

$$\sup_{\pi_k \in \Theta_k} \int_{\Omega_k} f(\mathbf{x}, \boldsymbol{\xi}) d\pi_k(\boldsymbol{\xi}) + z_k \left(1 - \int_{\Omega_k} d\pi_k(\boldsymbol{\xi})\right) + \mathbf{z}_k^T \left(\hat{\boldsymbol{\mu}}_k - \int_{\Omega_k} \boldsymbol{\xi} d\pi_k(\boldsymbol{\xi})\right)$$
$$+ \mathbf{Z}_k \circ \left[(\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^T) - \int_{\Omega_k} \boldsymbol{\xi}\boldsymbol{\xi}^T d\pi_k(\boldsymbol{\xi})\right] =$$
$$\sup_{\pi_k \in \Theta_k} z_k + \hat{\boldsymbol{\mu}}_k^T \mathbf{z}_k + (\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^T) \circ \mathbf{Z}_k + \int_{\Omega_k} \left[f(\mathbf{x}, \boldsymbol{\xi}) - z_k - \mathbf{z}_k^T \boldsymbol{\xi} - \boldsymbol{\xi}^T \mathbf{Z}_k \boldsymbol{\xi}\right] d\pi_k(\boldsymbol{\xi}).$$

This will be bounded above if and only if the integrand is nonpositive for every $\boldsymbol{\xi} \in \Omega_k$, and in this case, the supremum will occur when $\pi_k = 0$ everywhere. Thus, the dual problem is:

$$\begin{aligned} \min_{z_k, \mathbf{z}_k, \mathbf{Z}_k} \quad & z_k + {\mathbf{z}_k}^T \hat{\boldsymbol{\mu}}_k + \mathbf{Z}_k \circ (\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^T), \\ \text{subject to} \quad & \boldsymbol{\xi}^T \mathbf{Z}_k \boldsymbol{\xi} + \mathbf{z}_k^T \boldsymbol{\xi} + z_k \geq f(\mathbf{x}, \boldsymbol{\xi}), \quad \forall \boldsymbol{\xi} \in \Omega_k. \end{aligned} \tag{B.2}$$

Now we consider a vector function $g(\mathbf{x})$ where each component $g_k(\mathbf{x})$ is defined as the optimal value to (B.2) for each $k = 1, ..., r$ and $g_k(\mathbf{x}) = f(\mathbf{x}, \mathbf{m}_k)$ for $k = r+1, ..., K$. We can rewrite the overall problem (2.3) as follows:

$$\begin{aligned} \min_{\mathbf{x} \in \mathscr{X}} \quad & \sup \quad \sum_{k=1}^{K} p_k \, g_k(\mathbf{x}), \\ & \text{subject to} \quad \mathbf{p} \in \mathscr{P}, \end{aligned} \tag{B.3}$$

where $\mathscr{P} = \{\mathbf{p} : \mathbf{e}^T \mathbf{p} = 1, \ \mathbf{p} \geq 0, \ \sum_{k=1}^{K} \hat{p}_k \phi(\frac{p_k}{\hat{p}_k}) \leq \rho\}$. This is equivalent to minimizing β, where $\beta \geq \mathbf{p}^T \mathbf{g}(\mathbf{x})$ for all $\mathbf{p} \in \mathscr{P}$. This robust constraint can be reformulated according to Theorem 1 in Ben-Tal et al. (2013). They formulate the dual problem and show strong duality. According to the theorem, $\mathbf{x}$ satisfies such an infinite dimensional constraint if and only if there exists $\eta, \lambda \in \mathbb{R}$ such that $(\mathbf{x}, \eta, \lambda)$ satisfy $\beta \geq \eta + \rho\lambda + \lambda \sum_{k=1}^{K} \hat{p}_k \phi^\star \left(\frac{g_k(\mathbf{x}) - \eta}{\lambda}\right)$, $\lambda \geq 0$. Therefore, we can replace (B.3) with:

$$\begin{aligned} \min_{\mathbf{x} \in \mathscr{X}} \quad & \eta + \rho\lambda + \lambda \sum_{k=1}^{K} \hat{p}_k \phi^\star \left(\frac{g_k(\mathbf{x}) - \eta}{\lambda}\right), \\ & \lambda \geq 0, \end{aligned} \tag{B.4}$$

where $0\phi^\star(b/0) = 0$ if $b \leq 0$ and $0\phi^\star(b/0) = \infty$ if $b > 0$. We have an additional dual feasibility constraint to ensure that $\phi^\star$ is finite. Recall that $\phi^\star(s) = \sup_{t\geq 0}\{st - \phi(t)\}$. Thus, if $\lim_{t\to\infty} \frac{\phi(t)}{t} = \infty$, then for all $s \in \mathbb{R}$, $\phi^\star(s) < \infty$, and dual feasibility is guaranteed. This is the case, for example, with KL divergence, and we need not include any dual feasibility constraints. However, if $\lim_{t\to\infty} \frac{\phi(t)}{t} = \bar{s} < \infty$, then for any $s > \bar{s}$, $\phi^\star(s) = \sup_{t\geq 0}\{st - \phi(t)\} = \infty$. Hence, in order to ensure dual feasibility for such ϕ-divergence functions, we require that $g_k(\mathbf{x}) - \eta \leq \left(\lim_{t\to\infty} \frac{\phi(t)}{t}\right)\lambda$, for all $k = 1, ..., K$. This is the case, for example, with variation distance.

Under Assumption 2, $\phi^\star$ is monotone increasing, and so $\phi^\star(\min\, G(z_k, \mathbf{z}_k, \mathbf{Z}_k)) = \min \phi^\star(G(z_k, \mathbf{z}_k, \mathbf{Z}_k))$ for any function G. Thus, we can reformulate (2.3) as follows:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \eta + \rho\lambda + \lambda\sum_{k=1}^{r} \hat{p}_k\phi^\star\left(\frac{z_k + {\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k + \mathbf{Z}_k \circ (\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T) - \eta}{\lambda}\right) \\
& + \lambda\sum_{k=r+1}^{K} \hat{p}_k\phi^\star\left(\frac{f(\mathbf{x}, \mathbf{m}_k) - \eta}{\lambda}\right), \\
\text{subject to} \quad & \boldsymbol{\xi}^T\mathbf{Z}_k\boldsymbol{\xi} + \mathbf{z}_k^T\boldsymbol{\xi} + z_k \geq f(\mathbf{x}, \boldsymbol{\xi}), \quad \forall\boldsymbol{\xi} \in \Omega_k,\ \forall k = 1, ..., r, \\
& z_k + {\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k + \mathbf{Z}_k \circ (\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T) - \eta \leq \left(\lim_{t\to\infty} \frac{\phi(t)}{t}\right)\lambda,\ \forall k = 1, ..., r, \\
& f(\mathbf{x}, \mathbf{m}_k) - \eta \leq \left(\lim_{t\to\infty} \frac{\phi(t)}{t}\right)\lambda, \quad \forall k = r+1, ..., K, \\
& \lambda \geq 0.
\end{aligned}
\tag{B.5}
$$

Given that Assumption 1 is satisfied, we get the following minimization problem:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{\gamma_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \eta+\rho\lambda+\lambda\sum_{k=1}^{r}\hat{p}_k\phi^\star\left(\frac{z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta}{\lambda}\right)\\
& +\lambda\sum_{k=r+1}^{K}\hat{p}_k\phi^\star\left(\frac{\gamma_k-\eta}{\lambda}\right),\\
\text{subject to} \quad & \begin{pmatrix}1 & \boldsymbol{\xi}^T\end{pmatrix}\begin{pmatrix}z_k-c_l-\boldsymbol{\beta}_l^T\mathbf{x} & \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x})^T\\ \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x}) & \mathbf{Z}_k-\mathbf{A}_l\end{pmatrix}\begin{pmatrix}1\\ \boldsymbol{\xi}\end{pmatrix}\geq 0\\
& \forall\boldsymbol{\xi}\in\Omega_k,\forall l=1,...,L,\ \forall k=1,...,r,\\
& z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\\
& \forall k=1,...,r,\\
& \gamma_k\geq(\boldsymbol{\beta}_l+\mathbf{D}_l\mathbf{m}_k)^T\mathbf{x}+\mathbf{m}_k^T\mathbf{A}_l\mathbf{m}_k+\mathbf{b}_l^T\mathbf{m}_k+c_l,\quad\forall\, l=1,...,L,\\
& \forall k=r+1,...,K,\\
& \gamma_k-\eta\leq\left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\quad\forall k=r+1,...,K,\\
& \lambda\geq 0.
\end{aligned}
\tag{B.6}
$$

As shown by Isii (1960), strong duality holds (the solution to (B.6) is equal to that of (2.2)), as long as $\hat{\boldsymbol{\Sigma}}_k\succ 0$ for all $k=1,...,K$. This condition holds with probability 1 for sample estimates of the covariance matrices from a continuous distribution, as there will always be at least two distinct points in Ω_k (if not, then $k\in S$, Ω_k is a point mass and no estimate is needed). These points will share no coordinate values with probability 1. Thus, if we compute $\hat{\boldsymbol{\Sigma}}_k=\frac{1}{N_k-1}\sum_{\boldsymbol{\xi}_i\in\Omega_k}(\boldsymbol{\xi}_i-\hat{\boldsymbol{\mu}}_k)(\boldsymbol{\xi}_i-\hat{\boldsymbol{\mu}}_k)^T$, where $N_k>1$ is the number of data points contained in region Ω_k, then $\mathbf{x}^T\hat{\boldsymbol{\Sigma}}_k\mathbf{x}=\frac{1}{N_k-1}\sum_{\boldsymbol{\xi}_i\in\Omega_k}\left[(\boldsymbol{\xi}_i-\hat{\boldsymbol{\mu}}_k)^T\mathbf{x}\right]^T\left[(\boldsymbol{\xi}_i-\hat{\boldsymbol{\mu}}_k)^T\mathbf{x}\right]>0$ for any $\mathbf{x}\neq 0$.

Lemma 7. *Problem* (2.4) *is convex for any* ϕ*-divergence.*

Proof. For any ϕ, $\phi^\star$ is convex since it is a supremum over linear functions. Let $\lambda\phi^\star(\frac{\mathscr{G}-\eta}{\lambda})=\sup_{t\geq 0}\{\mathscr{G}-\eta t-\lambda\phi(t)\}$ where $\mathscr{G}$ is a linear function of any optimization variables $\mathbf{x}$, λ, and/or the dual variables. Since the supremum over linear functions

is convex, $\lambda\phi^\star(\frac{\mathscr{G}-\eta}{\lambda})$ is jointly convex in any of these variables and thus the objective function in (2.4) is convex. Clearly, the infinite dimensional constraints can be reformulated as an infimum over linear functions which is concave, and thus the constraints are convex. □

Suppose we embed our objective function in the constraints such as

$$\eta+\rho\lambda+\lambda\sum_{k=1}^{r}\hat{p}_k\phi^\star\left(\frac{z_k+\mathbf{z}_k{}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\cdot(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)-\eta}{\lambda}\right)+\lambda\sum_{k=r+1}^{K}\hat{p}_k\phi^\star\left(\frac{\gamma_k-\eta}{\lambda}\right)\leq t.$$

For many ϕ, this constraint set admits a self-concordant barrier function, as shown in Pardo (2006) for the Burg entropy-based divergence and Kullback-Leibler divergence. We will denote such a class of ϕ-divergence functions as Φ and restrict our consideration to this class.

In the special case where there is no probability ambiguity ($\rho = 0$), our problem looks like:

$$\begin{aligned}\min_{\mathbf{x}\in\mathscr{X}}\ & \sum_{k\in C}\hat{p}_k\sup_{\pi_k\in\Theta_k}\int_{\Omega_k}f(\mathbf{x},\boldsymbol{\xi})d\pi_k(\boldsymbol{\xi})+\sum_{k\in S}\hat{p}_k f(\mathbf{x},\mathbf{m}_k),\\ \text{subject to}\ & \int_{\Omega_k}d\pi_k(\boldsymbol{\xi})=1,\\ & \int_{\Omega_k}\boldsymbol{\xi}d\pi_k(\boldsymbol{\xi})=\hat{\boldsymbol{\mu}}_k,\\ & \int_{\Omega_k}\boldsymbol{\xi}\boldsymbol{\xi}^T d\pi_k(\boldsymbol{\xi})=\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T,\ \forall k=1,...,r,\end{aligned}\tag{B.7}$$

We may simply combine the dual of each inner moment problem as in (B.2) as follows:

$$\begin{aligned}\min_{\mathbf{x}\in\mathscr{X},z_k,\mathbf{z}_k,\mathbf{Z}_k}\ & \sum_{k=1}^{r}\hat{p}_k\left(z_k+\mathbf{z}_k{}^T\mu_k+\mathbf{Z}_k\circ(\boldsymbol{\Sigma}_k+\boldsymbol{\mu}_k\boldsymbol{\mu}_k^T)\right)+\sum_{k=r+1}^{K}\hat{p}_k f(\mathbf{x},\mathbf{m}_k),\\ \text{subject to}\ & \boldsymbol{\xi}^T\mathbf{Z}_k\boldsymbol{\xi}+\mathbf{z}_k^T\boldsymbol{\xi}+z_k\geq f(\mathbf{x},\boldsymbol{\xi}),\quad\forall\boldsymbol{\xi}\in\Omega_k,\ \forall k=1,...,r.\end{aligned}\tag{B.8}$$

Given that Assumption 1 is satisfied, we get:

$$\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X},\{\gamma_k,z_k,\mathbf{z}_k,\mathbf{Z}_k\}} \quad & \sum_{k=1}^{r}\hat{p}_k\left(z_k+{\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k+\mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k+\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T)\right)+\sum_{k=r+1}^{K}\hat{p}_k\gamma_k,\\
\text{subject to} \quad & \begin{pmatrix}1 & \boldsymbol{\xi}^T\end{pmatrix}\begin{pmatrix} z_k-c_l-\boldsymbol{\beta}_l^T\mathbf{x} & \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x})^T\\ \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x}) & \mathbf{Z}_k-\mathbf{A}_l\end{pmatrix}\begin{pmatrix}1\\ \boldsymbol{\xi}\end{pmatrix}\geq 0,\\
& \forall\boldsymbol{\xi}\in\Omega_k,\forall l=1,...,L,\ \forall k=1,...,r,\\
& \gamma_k\geq(\boldsymbol{\beta}_l+\mathbf{D}_l\mathbf{m}_k)^T\mathbf{x}+\mathbf{m}_k^T\mathbf{A}_l\mathbf{m}_k+\mathbf{b}_l^T\mathbf{m}_k+c_l,\quad \forall\, l=1,...,L,\\
& \forall k=r+1,...,K.
\end{aligned}\tag{B.9}$$

□

B.2 Proof of Lemma 1

Proof. Let $\bar{\mathbf{Z}}_{kl}=\begin{pmatrix} z_k-c_l-\boldsymbol{\beta}_l^T\mathbf{x} & \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x})^T\\ \frac{1}{2}(\mathbf{z}_k-\mathbf{b}_l-\mathbf{D}_l^T\mathbf{x}) & \mathbf{Z}_k-\mathbf{A}_l\end{pmatrix}$. We are concerned with the mr infinite dimensional constraints:

$$\begin{pmatrix}1 & \boldsymbol{\xi}^T\end{pmatrix}\bar{\mathbf{Z}}_{kl}\begin{pmatrix}1\\ \boldsymbol{\xi}\end{pmatrix}\geq 0,\quad \forall\boldsymbol{\xi}\in\Omega_k,\quad \text{for } l=1,...,L,\ \ k=1,...,r. \tag{B.10}$$

Let $\bar{\boldsymbol{\xi}}=\begin{pmatrix}1\\ \boldsymbol{\xi}\end{pmatrix}$. Note that the constraint of containment in a polytope Ω_k can be represented as $\mathbf{A}_k\bar{\boldsymbol{\xi}}\leq\mathbf{b}_k$ for appropriate $\mathbf{A}_k$ and $\mathbf{b}_k$. By adding slack variables and constructing a higher dimensional vector $\hat{\boldsymbol{\xi}}$, this is equivalent to $\hat{\mathbf{A}}_k\hat{\boldsymbol{\xi}}=\hat{\mathbf{b}}_k$. Some of the slack variables may be required to be nonnegative. This can be encoded as $\hat{\boldsymbol{\xi}}\in\mathcal{C}$ where $\mathcal{C}$ is a cone where some elements are required to be nonnegative while others are free. If any have lower and upper bounds, we may simply introduce two more slack variables to represent the distance to each bound, both of which are required to be nonnegative.

For notational convenience, we will drop the indices l and k and consider a generic constraint with matrix $\bar{\mathbf{Z}}$ and region Ω described by the linear constraints $\mathbf{A}\hat{\boldsymbol{\xi}}=\mathbf{b}$. Let $\boldsymbol{\Xi}=\hat{\boldsymbol{\xi}}(\hat{\boldsymbol{\xi}})^T$. Let $\bar{p}$ be the dimension of $\hat{\boldsymbol{\xi}}$, and addend to $\bar{\mathbf{Z}}$ rows and columns of zeros

so that it is a $\bar{p}$-by-$\bar{p}$ matrix (it is originally $(p+1)$-by-$(p+1)$). Call this new matrix $\hat{\mathbf{Z}}$. Consider the minimization problem

$$\begin{aligned} \min_{\boldsymbol{\Xi}} \ & \langle \hat{\mathbf{Z}}, \boldsymbol{\Xi} \rangle, \\ \text{subject to } & \mathbf{A}\hat{\boldsymbol{\xi}} = \mathbf{b}, \\ & \hat{\boldsymbol{\xi}} \in \mathcal{C}, \\ & \boldsymbol{\Xi} = \hat{\boldsymbol{\xi}}(\hat{\boldsymbol{\xi}})^T. \end{aligned} \tag{B.11}$$

For constraint (B.10) to hold is equivalent to the objective value for this problem to be nonnegative. Since the objective function is linear, we can take the closure of the convex hull of the feasible set and the solution will remain the same. By Theorem 1 from Burer and Dong (2012), the closure of the convex hull can be represented as

$$\left\{ (\hat{\boldsymbol{\xi}}, \boldsymbol{\Xi}) \mid \begin{pmatrix} 1 & \hat{\boldsymbol{\xi}}^T \\ \hat{\boldsymbol{\xi}} & \boldsymbol{\Xi} \end{pmatrix} \in \mathcal{CP}(\mathbb{R}^+ \times \mathcal{C}),\ \mathbf{A}\hat{\boldsymbol{\xi}} = \mathbf{b},\ (\mathbf{A}\boldsymbol{\Xi}\mathbf{A}^T)_{ii} = b_i^2 \ \ \forall i \right\}, \tag{B.12}$$

where $\mathcal{CP}(\mathcal{K}) = \text{cl conv } \{xx^T | x \in \mathcal{K}\}$ is the set of generalized completely positive matrices over a cone $\mathcal{K}$. The constraint defining $\boldsymbol{\Xi}$ in terms of $\hat{\boldsymbol{\xi}}$ is now freed, although they are still connected by the completely copositive constraint. $\hat{\mathbf{Z}}$ need simply satisfy that

$$\left\langle \begin{pmatrix} 0 & 0 \\ 0 & \hat{\mathbf{Z}} \end{pmatrix}, \begin{pmatrix} 1 & \hat{\boldsymbol{\xi}}^T \\ \hat{\boldsymbol{\xi}} & \boldsymbol{\Xi} \end{pmatrix} \right\rangle \geq 0 \text{ for all } \hat{\boldsymbol{\xi}}, \boldsymbol{\Xi} \text{ that satisfy (B.12).}$$

The two latter sets of constraints in (B.12) are simply affine constraints, and thus we can represent these constraints as

$$\hat{\boldsymbol{\Xi}} = \begin{pmatrix} 1 & \hat{\boldsymbol{\xi}}^T \\ \hat{\boldsymbol{\xi}} & \boldsymbol{\Xi} \end{pmatrix} \in \mathcal{A} \text{ for some affine space } \mathcal{A}.$$

In $\mathcal{C}$, some of the elements are required to be nonnegative and others free, so the same is true for $\mathbb{R}^+ \times \mathcal{C}$. Thus, $\mathcal{CP}(\mathbb{R}^+ \times \mathcal{C})$ is the cone of completely positive matrices over such a cone, which we call a "mixed" cone. Let

$$\hat{\hat{\mathbf{Z}}} = \begin{pmatrix} 0 & 0 \\ 0 & \hat{\mathbf{Z}} \end{pmatrix}.$$

We have that $\langle \hat{\hat{\mathbf{Z}}}, \hat{\boldsymbol{\Xi}} \rangle \geq 0$ for all $\hat{\boldsymbol{\Xi}}$ such that $\hat{\boldsymbol{\Xi}} \in \mathcal{A} \cap \mathcal{CP}(\mathbb{R}^+ \times \mathcal{C})$. This is true iff $\langle \hat{\hat{\mathbf{Z}}}, \hat{\boldsymbol{\Xi}} \rangle \geq 0$ holds for all $\hat{\boldsymbol{\Xi}} \in \text{conic}[\mathcal{A} \cap \mathcal{CP}(\mathbb{R}^+ \times \mathcal{C})]$ as points in the conic hull are linear combinations with positive coefficients. Thus, we need $\hat{\hat{\mathbf{Z}}}$ to be in the dual of this cone. The dual cone of a given cone $\mathcal{C} \subseteq \mathcal{S}$ is given by $\mathcal{C}^\star = \{\mathbf{A} \in \mathcal{S} : \langle \mathbf{A}, \mathbf{B} \rangle \geq 0 \; \forall \, \mathbf{B} \in \mathcal{C}\}$. In particular, $\mathcal{C}_n^\star = \mathcal{CP}_n$. The conic hull will be an intersection of the cone with a linear subspace $\mathcal{L}$. The dual cone of a mixed cone is a mixed copositive cone, which we denote by $\mathcal{CO}(\mathbb{R}^+ \times \mathcal{C})$. The dual of $\mathcal{L} \cap \mathcal{CP}(\mathbb{R}^+ \times \mathcal{C})$ is given by $\mathcal{L}^\perp + \mathcal{CO}(\mathbb{R}^+ \times \mathcal{C})$. Thus, our constraint is that $\hat{\hat{\mathbf{Z}}} \in \mathcal{L}^\perp + \mathcal{CO}(\mathbb{R}^+ \times \mathcal{C})$.

If we add back in the indices l and k, we replace the original copositive constraints with $\hat{\hat{\mathbf{Z}}}_{kl} \in \mathcal{L}_k^\perp + \mathcal{CO}(\mathbb{R}^+ \times \mathcal{C}_k)$. Hence, instead of (2.8), we can solve:

$$
\begin{aligned}
\min_{\mathbf{x} \in \mathscr{X}, \eta, \lambda, \{\gamma_k, z_k, \mathbf{z}_k, \mathbf{Z}_k\}} \quad & \eta + \rho\lambda + \lambda \sum_{k=1}^{r} \hat{p}_k \phi^\star \left(\frac{z_k + {\mathbf{z}_k}^T \hat{\boldsymbol{\mu}}_k + \mathbf{Z}_k \circ (\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^T) - \eta}{\lambda} \right) \\
& + \lambda \sum_{k=r+1}^{K} \hat{p}_k \phi^\star \left(\frac{\gamma_k - \eta}{\lambda} \right), \\
\text{subject to} \quad & \hat{\hat{\mathbf{Z}}}_{kl} \in \mathcal{L}_k^\perp + \mathcal{CO}(\mathbb{R}^+ \times \mathcal{C}_k), \\
& \forall \; l = 1, ..., L, \; \forall k = 1, ..., r, \\
& z_k + {\mathbf{z}_k}^T \hat{\boldsymbol{\mu}}_k + \mathbf{Z}_k \circ (\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^T) - \eta \leq \left(\lim_{t \to \infty} \frac{\phi(t)}{t} \right) \lambda, \\
& \forall k = 1, ..., r, \\
& \gamma_k \geq (\boldsymbol{\beta}_l + \mathbf{D}_l \mathbf{m}_k)^T \mathbf{x} + \mathbf{m}_k^T \mathbf{A}_l \mathbf{m}_k + \mathbf{b}_l^T \mathbf{m}_k + c_l, \; \forall \; l = 1, ..., L, \\
& \forall k = r+1, ..., K, \\
& \gamma_k - \eta \leq \left(\lim_{t \to \infty} \frac{\phi(t)}{t} \right) \lambda, \; \forall k = r+1, ..., K, \\
& \lambda \geq 0.
\end{aligned}
\tag{B.13}
$$

When we construct $\hat{\boldsymbol{\xi}}$ constrained by $\hat{\mathbf{A}}_k \hat{\boldsymbol{\xi}} = \hat{\mathbf{b}}_k$, the increase in dimension over $\bar{\boldsymbol{\xi}}$ cannot exceed the number of constraints in the linear inequality representation of Ω_k, $\mathbf{A}_k \bar{\boldsymbol{\xi}} \leq \mathbf{b}_k$, equal to the number of sides of the polytope. This is because we may only introduce one slack variable for each inequality. For those slack variables that are constrained

on both sides, we introduce two more slack variables for a total increase bounded by twice the number of slack variables. Thus, the total size of the constraint only grows polynomially.

□

B.3 Proof of Theorem 5

Proof. Consider the conditional moments problem, and suppose $\pi^\star$ is the true governing distribution for $\boldsymbol{\xi}$. Suppose the cost function $f(\mathbf{x}, \boldsymbol{\xi})$ is bounded by M and Lipschitz in $\boldsymbol{\xi}$ with Lipschitz constant C. Suppose the region Ω is compact and we partition into r subregions $\{\Omega_k\}_r$, all with equal diameter d. Then, for any fixed $\mathbf{x}$, the absolute difference between the objective function values $v_N(\mathbf{x})$ and $v^\star(\mathbf{x})$ can be bounded as

below:

$$
\begin{aligned}
&|v_N(\mathbf{x}) - v^\star(\mathbf{x})| \\
&= \left| \sup_{\pi \in \Theta} \mathbf{E}_\pi[f(\mathbf{x}, \tilde{\boldsymbol{\xi}})] - \mathbf{E}_{\pi^\star}[f(\mathbf{x}, \tilde{\boldsymbol{\xi}})] \right| \\
&= \left| \sup_{\mathbf{p} \in \mathscr{P}} \sum_{k=1}^{r} p_k \sup_{\pi_k} \mathbf{E}_{\pi_k}[f(\mathbf{x}, \tilde{\boldsymbol{\xi}})] - \sum_{k=1}^{r} p_k^\star \mathbf{E}_{\pi_k^\star}[f(\mathbf{x}, \tilde{\boldsymbol{\xi}})] \right| \\
&= \left| \sup_{\mathbf{p} \in \mathscr{P}} \left\{ \sum_{k=1}^{r} \sup_{\pi_k} \mathbf{E}_{\pi_k}[f(\mathbf{x}, \tilde{\boldsymbol{\xi}})](p_k - p_k^\star) \right\} + \sum_{k=1}^{r} p_k^\star \left(\sup_{\pi_k} \mathbf{E}_{\pi_k}[f(\mathbf{x}, \tilde{\boldsymbol{\xi}})] - \mathbf{E}_{\pi_k^\star}[f(\mathbf{x}, \tilde{\boldsymbol{\xi}})] \right) \right| \\
&\leq \left| \sup_{\mathbf{p} \in \mathscr{P}} \left\{ \sum_{k=1}^{r} \sup_{\pi_k} \mathbf{E}_{\pi_k}[f(\mathbf{x}, \tilde{\boldsymbol{\xi}})](p_k - \hat{p}_k) \right\} + \sum_{k=1}^{r} \sup_{\pi_k} \mathbf{E}_{\pi_k}[f(\mathbf{x}, \tilde{\boldsymbol{\xi}})](\hat{p}_k - p_k^\star) \right| \\
&+ \sum_{k=1}^{r} p_k^\star \left| \sup_{\pi_k} \mathbf{E}_{\pi_k}[f(\mathbf{x}, \tilde{\boldsymbol{\xi}})] - \mathbf{E}_{\pi_k^\star}[f(\mathbf{x}, \tilde{\boldsymbol{\xi}})] \right| \\
&\leq \left| \sup_{\mathbf{p} \in \mathscr{P}} \sum_{k=1}^{r} \sup_{\Omega_k} |f(\mathbf{x}, \boldsymbol{\xi})|(p_k - \hat{p}_k) \right| + \sum_{k=1}^{r} \sup_{\Omega_k} |f(\mathbf{x}, \boldsymbol{\xi})||\hat{p}_k - p_k^\star| \\
&+ \sum_{k=1}^{r} p_k^\star \sup_{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \Omega_k} |f(\mathbf{x}, \boldsymbol{\xi}_1) - f(\mathbf{x}, \boldsymbol{\xi}_2)| \\
&\leq \sup_{\Omega} |f(\mathbf{x}, \boldsymbol{\xi})| \sup_{\mathbf{p} \in \mathscr{P}} \sum_{k=1}^{r} |p_k - \hat{p}_k| + \sup_{\Omega} |f(\mathbf{x}, \boldsymbol{\xi})| \sum_{k=1}^{r} |\hat{p}_k - p_k^\star| + Cd \\
&\leq M \sup_{\mathbf{p} \in \mathscr{P}} \sum_{k=1}^{r} |p_k - \hat{p}_k| + M \sum_{k=1}^{r} |\hat{p}_k - p_k^\star| + Cd.
\end{aligned}
$$

Note that this bound is independent of $\mathbf{x}$, and thus holds for $|\Upsilon_N - \Upsilon^\star| = |\min_{\mathbf{x}} v_N(\mathbf{x}) - \min_{\mathbf{x}} v^\star(\mathbf{x})|$. We know that $d \to 0$ as the number of regions $r \to \infty$ since Ω is compact. If we denote the dimension of $\boldsymbol{\xi}$ by p, then d is proportional to $\frac{1}{r^{\frac{1}{p}}}$. Thus the third term goes to 0, but we have an infinite number of terms in the sums in the first two terms. We can view each $\hat{p}_k$ as a realization of a binomial random variable, divided by the number of trials N, with success probability $p_k^\star$. Thus the difference between the sample mean $\hat{p}_k$ and the true mean $p_k^\star$ converges almost surely to zero under the strong law of large numbers. By Chebyshev's Inequality, $\mathcal{P}\left(|\hat{p}_k - p_k^\star| \geq \frac{a}{\sqrt{N}} \sqrt{p_k^\star(1 - p_k^\star)}\right) \leq \frac{1}{a^2}$. We can solve for the supremum of $\sum_k \sqrt{p_k^\star(1 - p_k^\star)}$ such that $\sum_k p_k^\star = 1$, which is given by $\sqrt{r-1}$, achieved when $p_k^\star = \frac{1}{r}$ for every $k = 1, ..., r$. Thus we can replace the

bound with:

$$\leq M \sup_{\mathbf{p}\in\mathscr{P}} \sum_{k=1}^{r} |p_k - \hat{p}_k| + M\frac{a}{\sqrt{N}}\sqrt{r-1} + C' r^{-\frac{1}{D}},$$

which holds with probability greater than or equal to $\frac{1}{a^2}$.

Lemma 8. *For many ϕ-divergence functions, $\sup_{\mathbf{p}\in\mathscr{P}} \sum_{k=1}^{r} |p_k - \hat{p}_k| \to 0$ as $N \to \infty$.*

Proof. Observe that $\mathscr{P} = \{\mathbf{p} : \mathbf{e}^T\mathbf{p} = 1,\ \mathbf{p} \geq 0,\ \sum_{k=1}^{K} \hat{p}_k \phi(\frac{p_k}{\hat{p}_k}) \leq \rho\}$ with $\rho = \frac{\phi''(1)}{2N}\chi^2_{K-1,1-\beta}$. Thus, as $N >> r \to \infty$, $\rho \to 0$, and thus $\frac{p_k}{\hat{p}_k} \to 1$ for every k (since $\phi \geq 0$ is convex and $\phi(1) = 0$). □

Suppose $\epsilon > 0$ and $\delta > 0$. Then we let a be such that $\frac{1}{a^2} = \delta$. Then we can choose r and N large enough ($N \gg r$) such that $M\frac{a}{\sqrt{N}}\sqrt{r-1} + C' r^{-\frac{1}{D}} \leq \frac{\epsilon}{2}$. By Lemma 8, we can choose N large enough such that $M \sup_{\mathbf{p}\in\mathscr{P}} \sum_{k=1}^{r} |p_k - \hat{p}_k| \leq \frac{\epsilon}{2}$. Choose N, r large enough such that both of these conditions hold. Then we have that $|\Upsilon_N - \Upsilon^\star| \leq \epsilon$ with probability $p \geq 1 - \delta$. Thus, $\mathcal{P}(|\Upsilon_N - \Upsilon^\star| \geq \epsilon) \leq \delta$, and $\mathcal{P}(|\Upsilon_N - \Upsilon^\star| \geq \epsilon) \to 0$ $\forall \epsilon > 0$. □

Proposition 5b):

Proof. For large N, we can use the central limit theorem approximation to get that $|p_k - p_k^\star| \leq \frac{\Phi_{1-\frac{\alpha}{4}}}{\sqrt{N}}\sqrt{p_k^\star(1-p_k^\star)}$ with probability $1 - \frac{\alpha}{2}$, $\forall k$. Thus we can replace the middle term in the bound with:

$$\leq M\frac{\Phi_{1-\frac{\alpha}{2}}}{\sqrt{N}}\sqrt{r-1},$$

which holds with probability $1 - \frac{\alpha}{2}$ asymptotically.

Lemma 9. *For many ϕ-divergence functions, such as variation distance, there exists a function $\gamma(r, N)$ s.t. $\sup_{\mathbf{p}\in\mathscr{P}} \sum_{k=1}^{r} |p_k - \hat{p}_k| \leq \gamma(r, N)$.*

Proof. Observe that $\mathscr{P} = \{\mathbf{p} : \mathbf{e}^T\mathbf{p} = 1,\ \mathbf{p} \geq 0,\ \sum_{k=1}^{K} \hat{p}_k \phi(\frac{p_k}{\hat{p}_k}) \leq \rho\}$ with $\rho = \frac{\phi''(1)}{2N}\chi^2_{r-1,1-\beta}$. The proof will be dependent upon the choice of ϕ-divergence function.

We only give a couple of examples here.
Variation distance ($\sum_{k=1}^{K} \hat{p}_k \phi(\frac{p_k}{\hat{p}_k}) = \sum_{k=1}^{r} |p_k - \hat{p}_k|$): In this case, it is trivial and $\gamma(r, N) = \frac{\phi''(1)}{2N} \chi^2_{r-1,1-\beta}$.
Modified χ^2 Distance ($\sum_{k=1}^{K} \hat{p}_k \phi(\frac{p_k}{\hat{p}_k}) = \sum_{k=1}^{r} \frac{(p_k - \hat{p}_k)^2}{\hat{p}_k}$): Note that $(p_k - \hat{p}_k)^2 \leq \hat{p}_k \rho$ $\forall k$ since all terms in the sum are positive. Thus, $|p_k - \hat{p}_k| = \sqrt{(p_k - \hat{p}_k)^2} \leq \sqrt{\rho \hat{p}_k}$ $\forall k$, and $\sum_{k=1}^{r} |p_k - \hat{p}_k| \leq \sqrt{\rho} \sum_{k=1}^{r} \sqrt{\hat{p}_k} \leq \sqrt{\rho} \sum_{k=1}^{r} \sqrt{\frac{1}{r}} = \sqrt{\rho}\sqrt{r}$. We let $\gamma(r, N) = \frac{\sqrt{\phi''(1)}\sqrt{r}}{\sqrt{2N}} \sqrt{\chi^2_{r-1,1-\beta}}$. □

Thus, in these cases, our bound becomes:

$$\leq M \left(\gamma(r, N) + \frac{\Phi_{1-\frac{\alpha}{2}}}{\sqrt{N}} \sqrt{r-1} \right) + C' r^{-\frac{1}{p}},$$

which holds asymptotically with probability $\geq 1 - \alpha$. □

B.4 Proof of Theorem 7

We can rewrite (6.1) as:

$$\min_x \ (h+b)\hat{p}_1 \sup_{\pi_1 \in \Theta_1} \{\mathbb{E}_{\pi_1}[(\Xi - x)^+]\} + (h+b)\hat{p}_2 \sup_{\pi_2 \in \Theta_2} \{\mathbb{E}_{\pi_2}[(\Xi - x)^+]\} - h(\hat{\mu} - x),$$

$$\begin{aligned}
\text{subject to } & \mathbb{E}_{\pi_1}[\Xi] = \hat{\mu}_1, \\
& \mathbb{E}_{\pi_2}[\Xi] = \hat{\mu}_2, \\
& \mathbb{E}_{\pi_1}[\Xi^2] = \hat{\sigma}_1^2 + \hat{\mu}_1^2, \\
& \mathbb{E}_{\pi_2}[\Xi^2] = \hat{\sigma}_2^2 + \hat{\mu}_2^2,
\end{aligned} \tag{B.14}$$

where π_1, π_2, lie in the spaces of probability measures Θ_1, Θ_2, over the measurable spaces $((-\infty, \beta], \mathscr{F})$, $([\beta, \infty), \mathscr{F})$, respectively.

Consider the first subproblem and let $U = \beta - \Xi \geq 0$. The first subproblem becomes:

$$\begin{aligned} \sup_{\pi_1 \in \Theta_1} \quad & \mathbb{E}_{\pi_1}[(-U - (x-\beta))^+], \\ \text{subject to} \quad & \mathbb{E}_{\pi_1}[U] = \beta - \hat{\mu}_1, \\ & \mathbb{E}_{\pi_1}[U^2] = \hat{\sigma}_1^2 + (\beta - \hat{\mu}_1)^2. \end{aligned} \tag{B.15}$$

Suppose $x - \beta \geq 0$. Then $(-u - (x-\beta)) \leq 0$ for all $u \geq 0$ and the optimal value is 0.

Now suppose $x - \beta < 0$. The dual of (B.15) is given by:

$$\begin{aligned} \min_{z,\zeta,\eta} \quad & z + \zeta(\beta - \hat{\mu}_1) + \eta(\hat{\sigma}_1^2 + (\beta - \hat{\mu}_1)^2), \\ \text{subject to} \quad & z + \zeta u + \eta u^2 \geq (-u - (x-\beta))^+, \quad \forall u \geq 0. \end{aligned} \tag{B.16}$$

Let $g(u) = z + \zeta u + \eta u^2$. The function $g(u)$ is feasible if it is nonnegative and lies above the line $(-u - (x-\beta))$, which has a positive intercept and slope of negative 1, in the positive quadrant. Consider the optimal dual solution $(z^\star, \zeta^\star, \eta^\star)$ and associated function $g^\star$. Suppose that the constraint is not binding anywhere. That is, $z^\star + \zeta^\star u + \eta^\star u^2 > (-u-(x-\beta))^+, \quad \forall u \geq 0$. Then choose small $\epsilon > 0$ such that $z^\star + \zeta^\star u + \eta^\star u^2 - \epsilon \geq (-u-(x-\beta))^+, \quad \forall u \geq 0$, and consider $z^{'} = z^\star - \epsilon$. Then the value of the dual objective at $(z^{'}, \zeta^\star, \eta^\star)$ is less than that at $(z^\star, \zeta^\star, \eta^\star)$ and the point is still feasible, a contradiction. Therefore, we must have that the constraint is binding at some point $c \geq 0$.

Suppose that the constraint is not binding at any point $0 \leq u < \beta - x$. Then it must be binding at $u = c \geq \beta - x$. Since $c \geq \beta - x > 0$ and the constraint is binding at c, it must be that $g^\star(c) = 0$. As $g^\star \geq 0$ for $u \geq 0$, we have that the vertex of the parabola occurs at $g^\star(c) = 0$. Note that the parabola must be strictly concave up as it lies above the intercept $\beta - x > 0$. Hence, we can write $g^\star(u) = q(u-c)^2 = qu^2 - 2qcu + c^2$ for some $q, c > 0$, and thus $z^\star = c^2$, $\zeta^\star = -2qc$, and $\eta^\star = q$. The dual objective value would then be $\mathcal{D}(g) = c^2 - 2qc(\beta - \hat{\mu}_1) + q(\hat{\sigma}_1^2 + (\beta - \hat{\mu}_1)^2)$, and $\frac{\partial \mathcal{D}}{\partial q} = -2c(\beta - \hat{\mu}_1) + (\hat{\sigma}_1^2 + (\beta - \hat{\mu}_1)^2)$ is independent of q. We can change q while maintaining feasibility since $g^\star$ does not intersect the line $-u - (x - \beta)$ or the horizontal axis at any other point (the constraint is not binding at any other $u \geq 0$). In other words, $q(u-c)^2 > 0$ for all other $u \geq 0$ and so we can find small δ such that $(q - \delta)(u - c)^2 \geq 0$ for all other $u \geq 0$. Changing the concavity slightly will not affect the feasibility of g. If $\frac{\partial \mathcal{D}}{\partial q}$ is strictly positive or negative,

we may change q to decrease $\mathcal{D}$ and reach a contradiction since clearly $g^\star$ is not optimal for the dual problem. If the derivative is equal to zero, then we may decrease q until g intersects $-u-(x-\beta)$ at some point c, $0 \leq c < \beta - x$. Therefore, we may assume that the constraint is binding at some point c, $0 \leq c < \beta - x$.

Suppose that the constraint is binding at $u = c$, $0 < c < \beta - x$, so that $g^\star(u)$ intersects $-u-(x-\beta)$ before the line crosses the horizontal axis. Then $g^\star(u)$ must be tangent to $(-u-(x-\beta))$ at such an intersection point $0 < c < \beta - x$. Then we can write $g(u)-(-u-(x-\beta)) = a(u-c)^2$ for some constant $a \geq 0$. Our constraint becomes $a(u-c)^2+(-u-(x-\beta)) \geq 0$ for all $u \geq 0$. We use the same argument in the previous paragraph to show that the constraint $a(u-c)^2+(-u-(x-\beta)) \geq 0$ must be binding for $d > \beta - x$, so that $a(d-c)^2 = -(-d-(x-\beta))$ and $g(d) = 0$ (otherwise, we may change a, retaining feasibility, and decrease objective value). Let $u_0 = c + \frac{1}{2a}$ be the minimum of $g(u)$. The constraint must be binding at the minimum u_0 of $g(u)$, $g(u_0) = 0$, so that $d = u_0$. The value is $g(u_0) = a(c+\frac{1}{2a}-c)^2+(-c-\frac{1}{2a}-x+\beta) = -\frac{1}{4a}-c-x+\beta = 0$, which implies that $a = \frac{-1}{4(c+x-\beta)}$. Therefore, the dual optimal solution is given by $z = \frac{-c^2}{4(c+x-\beta)} - (x-\beta)$, $\zeta = \frac{2c}{4(c+x-\beta)} - 1$, and $\eta = \frac{-1}{4(c+x-\beta)}$. In this case, the dual objective is $\mathcal{D}(c) = \frac{-c^2}{4(c+x-\beta)} - (x-\beta) + (\frac{2c}{4(c+x-\beta)} - 1)(\beta - \hat{\mu}_1) - \frac{1}{4(c+x-\beta)}(\hat{\sigma}_1^2 + (\beta-\hat{\mu}_1)^2)$. We consider $\frac{d\mathcal{D}}{dc} = \frac{-4c^2-8c(x-\beta)}{16(c+x-\beta)} + \frac{8(x-\beta)(\beta-\hat{\mu}_1)}{16(c+x-\beta)} + \frac{4(\hat{\sigma}_1^2+(\beta-\hat{\mu}_1)^2)}{16(c+x-\beta)} = 0$. Solving this equation leads to $c^\star = -(x-\beta) - \sqrt{(x-\beta)^2 + 2(x-\beta)(\beta-\hat{\mu}_1) + (\hat{\sigma}_1^2 + (\beta-\hat{\mu}_1)^2)}$ (we take the negative root since $c^\star < \beta - x$). Thus, the worst case expected cost is $\mathcal{D}(c^\star) = -\frac{1}{2}((x-\beta)+(\beta-\hat{\mu}_1)) + \frac{1}{2}\sqrt{\hat{\sigma}_1^2 + ((x-\beta)+(\beta-\hat{\mu}_1))^2}$. Note that we rely on the fact that the point of tangency $c^\star > 0$, which is equivalent to $x - \beta \leq \frac{-(\beta-\hat{\mu}_1)^2-\hat{\sigma}_1^2}{2(\beta-\hat{\mu}_1)}$.

The remaining case to consider is when the constraint is binding at $u = 0$, thus there is no guaranteed tangency condition. We have that the intercept $z = -(x-\beta)$. It is always the case that the minimum of $g(u)$ occurs when $u^\star = -\frac{\zeta}{2\eta}$, as long as $\eta \neq 0$. If $\eta = 0$, then $\zeta \geq 0$ for the constraint $z + \zeta u + \eta u^2 \geq (-u-(x-\beta))^+$ to hold for all $u \geq 0$. But then we can choose $\eta > 0$ and the constraint will still hold, while increasing the dual objective since $\eta(\hat{\sigma}_1^2 + (\beta-\hat{\mu}_1)^2) > 0$. Therefore, $\eta > 0$. Since the value at the minimum must be $g(u^\star) = 0$, we have that $\eta = -\frac{\zeta^2}{4(x-\beta)}$. The dual objective is thus $\mathcal{D}(\zeta) = -(x-\beta) + \zeta(\beta-\hat{\mu}_1) - \frac{\zeta^2(\hat{\sigma}_1^2+(\beta-\hat{\mu}_1)^2)}{4(x-\beta)}$. We solve the first order optimality condition $\frac{d\mathcal{D}}{d\zeta} = (\beta-\hat{\mu}_1) - \frac{\zeta(\hat{\sigma}_1^2+(\beta-\hat{\mu}_1)^2)}{2(x-\beta)} = 0$ to arrive at $\zeta^\star = \frac{2(x-\beta)(\beta-\hat{\mu}_1)}{(\hat{\sigma}_1^2+(\beta-\hat{\mu}_1)^2)}$, and find

that $\mathcal{D}(\zeta^\star) = -(x-\beta) + \frac{(x-\beta)(\beta-\hat{\mu}_1)^2}{\hat{\sigma}_1^2+(\beta-\hat{\mu}_1)^2}$, which holds when $x - \beta \geq \frac{-(\beta-\hat{\mu}_1)^2-\hat{\sigma}_1^2}{2(\beta-\hat{\mu}_1)}$.

Consider the second subproblem and let $v = \xi - \beta \geq 0$. The second subproblem becomes:

$$\begin{aligned} \sup_{\pi_1 \in \Theta_1} \quad & \mathbb{E}_{\pi_2}[(v-(x-\beta))^+], \\ \text{subject to} \quad & \mathbb{E}_{\pi_2}[v] = \hat{\mu}_2 - \beta, \\ & \mathbb{E}_{\pi_2}[v^2] = \hat{\sigma}_2^2 + (\hat{\mu}_2 - \beta)^2. \end{aligned} \tag{B.17}$$

According to Theorem 3 of Bertsimas et al. (2002), the optimal value of (B.17) is given by:

$$\left\{ \begin{array}{ll} \frac{1}{2}\left[(\hat{\mu}_2-\beta-(x-\beta)) + \sqrt{\hat{\sigma}_2^2 + (\hat{\mu}_2-\beta-(x-\beta))^2}\right], & \text{if } x-\beta \geq \frac{(\hat{\mu}_2-\beta)^2+\hat{\sigma}_2^2}{2(\hat{\mu}_2-\beta)}, \\ \hat{\mu}_2 - \beta - (x-\beta) + (x-\beta)\frac{\hat{\sigma}_2^2}{(\hat{\mu}_2-\beta)^2+\hat{\sigma}_2^2}, & \text{if } 0 \leq x-\beta < \frac{(\hat{\mu}_2-\beta)^2+\hat{\sigma}_2^2}{2(\hat{\mu}_2-\beta)}. \end{array} \right\}$$

If $x - \beta < 0$, then $\mathbb{E}_{\pi_2}[(v-(x-\beta))^+] = \mathbb{E}_{\pi_2}[v-(x-\beta)] = \hat{\mu}_2 - x$.

Let $q = h + b$. If we combine the two objective functions as in (B.14) for each of the cases, we have the following outer layer optimization problem:

$$\min_x \left\{ \begin{array}{ll} \frac{1}{2}q\hat{p}_1\left[-(x-\hat{\mu}_1) + \sqrt{\hat{\sigma}_1^2+(x-\hat{\mu}_1)^2}\right] + q\hat{p}_2(\hat{\mu}_2 - x) - h(\hat{\mu}-x), & (i) \\ q\hat{p}_1\left[-(x-\beta) + \frac{(x-\beta)(\beta-\hat{\mu}_1)^2}{\hat{\sigma}_1^2+(\beta-\hat{\mu}_1)^2}\right] + q\hat{p}_2(\hat{\mu}_2 - x) - h(\hat{\mu}-x), & (ii) \\ q\hat{p}_2\left[(\hat{\mu}_2-x) + (x-\beta)\frac{\hat{\sigma}_2^2}{(\hat{\mu}_2-\beta)^2+\hat{\sigma}_2^2}\right] - h(\hat{\mu}-x), & (iii) \\ \frac{1}{2}q\hat{p}_2\left[(\hat{\mu}_2-x) + \sqrt{\hat{\sigma}_2^2+(\hat{\mu}_2-x)^2}\right] - h(\hat{\mu}-x), & (iv) \end{array} \right\},$$

where the conditions are as follows:

$$\begin{aligned} (i): \quad & x - \beta \leq \frac{-(\beta-\hat{\mu}_1)^2 - \hat{\sigma}_1^2}{2(\beta-\hat{\mu}_1)}, \\ (ii): \quad & \frac{-(\beta-\hat{\mu}_1)^2 - \hat{\sigma}_1^2}{2(\beta-\hat{\mu}_1)} \leq x - \beta \leq 0, \\ (iii): \quad & 0 \leq x - \beta \leq \frac{(\hat{\mu}_2-\beta)^2 + \hat{\sigma}_2^2}{2(\hat{\mu}_2-\beta)}, \\ (iv): \quad & \frac{(\hat{\mu}_2-\beta)^2 + \hat{\sigma}_2^2}{2(\hat{\mu}_2-\beta)} \leq x - \beta. \end{aligned}$$

We denote this objective function as $v(x)$. Let $\bar{r}_1 = \frac{2h}{(h+b)\hat{p}_2} - 1$, $\bar{r}_2 = \frac{2b}{(h+b)\hat{p}_1} - 1$. Note

that if $h = b = 1$ (equal backlog and holding costs, WLOG), then $\bar{r}_1 = \frac{\hat{p}_1}{\hat{p}_2}$ and $\bar{r}_2 = \frac{\hat{p}_2}{\hat{p}_1}$.

We define the following terms for notational convenience:

$\beta_1 = \beta - \frac{(\beta-\hat{\mu}_1)^2+\hat{\sigma}_1^2}{2(\beta-\hat{\mu}_1)}$,

$\beta_2 = \beta + \frac{(\hat{\mu}_2-\beta)^2+\hat{\sigma}_2^2}{2(\hat{\mu}_2-\beta)}$,

$\alpha = v\left(\hat{\mu}_1 + \frac{\bar{r}_2\sigma_1}{\sqrt{1-\bar{r}_2^2}}\right) = \frac{1}{2}(h+b)\hat{p}_1\frac{\hat{\sigma}_1(1+\bar{r}_2)}{\sqrt{1-\bar{r}_2^2}} - \frac{b\bar{r}_2\hat{\sigma}_1}{\sqrt{1-\bar{r}_2^2}} + (h+b)\hat{p}_2(\hat{\mu}_2-\hat{\mu}_1) - h(\hat{\mu}-\hat{\mu}_1)$,

$\gamma = v(\beta_1) = b\frac{(\beta-\hat{\mu}_1)^2+\hat{\sigma}_1^2}{2(\beta-\hat{\mu}_1)} - \frac{1}{2}(h+b)\hat{p}_1(\beta-\hat{\mu}_1) + (h+b)\hat{p}_2(\hat{\mu}_2-\beta) - h(\hat{\mu}-\beta)$,

$\zeta = v(\beta) = (h+b)\hat{p}_2(\hat{\mu}_2-\beta) - h(\hat{\mu}-\beta)$,

$\tau = v(\beta_2) = (\hat{\mu}_2-\beta)\left\{\frac{\hat{p}_2}{2}(h+b) + \frac{h}{2} + \frac{h\hat{\sigma}_2^2}{2(\hat{\mu}_2-\beta)^2}\right\} - h(\hat{\mu}-\beta)$,

$\omega = v\left(\hat{\mu}_2 - \frac{\bar{r}_1\sigma_2}{\sqrt{1-\bar{r}_1^2}}\right) = \frac{1}{2}(h+b)\hat{p}_2\frac{\hat{\sigma}_2(\bar{r}_1+1)}{\sqrt{1-\bar{r}_1^2}} - h(\hat{\mu}-\hat{\mu}_2) - \frac{h\bar{r}_1\hat{\sigma}_2}{\sqrt{1-\bar{r}_1^2}}$,

$\kappa = v'(x)_{\beta_1\le x\le\beta} = -b + (h+b)\hat{p}_1\frac{(\beta-\hat{\mu}_1)^2}{\hat{\sigma}_1^2+(\beta-\hat{\mu}_1)^2}$,

$\lambda = v'(x)_{\beta\le x\le\beta_2} = h - (h+b)\hat{p}_2\frac{(\hat{\mu}_2-\beta)^2}{(\hat{\mu}_2-\beta)^2+\hat{\sigma}_2^2}$.

If $\bar{r}_1 \ge 1$, define $\omega = \infty$, and if $\bar{r}_2 \ge 1$, define $\alpha = \infty$, and we assume that $\hat{\mu}_1 \ne \beta \ne \hat{\mu}_2$.

Then the solution, determined from elementary calculus, is given by:

$$x_\star = \left\{ \begin{array}{ll} \hat{\mu}_1 + \frac{\bar{r}_2\sigma_1}{\sqrt{1-\bar{r}_2^2}}, & \text{if (i)} \\ \beta - \frac{(\beta-\hat{\mu}_1)^2+\hat{\sigma}_1^2}{2(\beta-\hat{\mu}_1)}, & \text{if (ii)} \\ \beta, & \text{if (iii)} \\ \beta + \frac{(\hat{\mu}_2-\beta)^2+\hat{\sigma}_2^2}{2(\hat{\mu}_2-\beta)}, & \text{if (iv)} \\ \hat{\mu}_2 - \frac{\bar{r}_1\sigma_2}{\sqrt{1-\bar{r}_1^2}}, & \text{if (v)} \end{array} \right\},$$

where the conditions are as follows:

(i) $\hat{\mu}_1 + \frac{\bar{r}_2\sigma_1}{\sqrt{1-\bar{r}_2^2}} < \beta_1$, $\alpha < \zeta$, $\alpha < \tau$, and $\alpha < \omega$ or $\hat{\mu}_2 - \frac{\bar{r}_1\sigma_2}{\sqrt{1-\bar{r}_1^2}} < \beta_2$,

(ii) $\hat{\mu}_1 + \frac{\bar{r}_2\sigma_1}{\sqrt{1-\bar{r}_2^2}} > \beta_1$, $\kappa > 0$, $\gamma < \tau$, and $\gamma < \omega$ or $\hat{\mu}_2 - \frac{\bar{r}_1\sigma_2}{\sqrt{1-\bar{r}_1^2}} < \beta_2$,

(iii) $\kappa < 0$, $\lambda > 0$, $\zeta < \alpha$ or $\hat{\mu}_1 + \frac{\bar{r}_2\sigma_1}{\sqrt{1-\bar{r}_2^2}} > \beta_1$, and $\zeta < \omega$ or $\hat{\mu}_2 - \frac{\bar{r}_1\sigma_2}{\sqrt{1-\bar{r}_1^2}} < \beta_2$,

(iv) $\lambda < 0$, $\tau < \gamma$, $\tau < \alpha$ or $\hat{\mu}_1 + \frac{\bar{r}_2\sigma_1}{\sqrt{1-\bar{r}_2^2}} > \beta_1$, and $\hat{\mu}_2 - \frac{\bar{r}_1\sigma_2}{\sqrt{1-\bar{r}_1^2}} < \beta_2$,

(v) $\hat{\mu}_2 - \frac{\bar{r}_1\sigma_2}{\sqrt{1-\bar{r}_1^2}} > \beta_2$, $\omega < \gamma$, $\omega < \zeta$, $\omega < \alpha$ or $\hat{\mu}_1 + \frac{\bar{r}_2\sigma_1}{\sqrt{1-\bar{r}_2^2}} > \beta_1$.

B.5 Proof of Theorem 10

Proof. In (8.5), the second term in the objective function can be broken up as:

$$\sup_{\mathbf{p}\in\mathscr{P}} \quad \sum_{k\in C} p_k \sup_{\pi_k\in\Theta_k} \int_{\Omega_k} Q(\mathbf{x},\boldsymbol{\xi}) d\pi_k(\boldsymbol{\xi}) + \sum_{k\in S} p_k Q(\mathbf{x},\mathbf{m}_k),$$

when the fragmentation consists of regions Ω_k and point masses $\mathbf{m}_k$, as before. The summation term does not involve the inner infinite dimensional variables, π_k. For each integral term, we can compute the dual of the kth inner moment problem independently as done in single stage:

$$\begin{aligned} \min_{z_k,\mathbf{z}_k,\mathbf{Z}_k} \quad & z_k + {\mathbf{z}_k}^T \hat{\boldsymbol{\mu}}_k + \mathbf{Z}_k \circ (\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^T), \\ \text{subject to} \quad & \boldsymbol{\xi}^T \mathbf{Z}_k \boldsymbol{\xi} + \mathbf{z}_k^T \boldsymbol{\xi} + z_k \geq Q(\mathbf{x},\boldsymbol{\xi}), \quad \forall \boldsymbol{\xi} \in \Omega_k. \end{aligned} \tag{B.18}$$

For notational convenience, in order to differentiate between the right hand side of the constraint for each k, we will consider each kth second stage problem with decision variable $\mathbf{w}_k$ as follows:

$$\begin{aligned} Q(\mathbf{x},\boldsymbol{\xi}) = \quad \min_{\mathbf{w}_k} \quad & \boldsymbol{\xi}^T \mathbf{w}_k, \\ \text{subject to} \quad & \mathbf{W}\mathbf{w}_k = \mathbf{h} - \mathbf{T}\mathbf{x}, \\ & \mathbf{w}_k \geq 0. \end{aligned} \tag{B.19}$$

Now we consider a vector function $g(\mathbf{x})$ where each component $g_k(\mathbf{x})$ is defined as the optimal value to (B.18) for each $k = 1, ..., r$, and $g_k(\mathbf{x}) = Q(\mathbf{x}, \mathbf{m}_k)$ for $k = r+1, ..., K$. We can rewrite the overall problem (8.5) as follows:

$$\begin{aligned} \min_{\mathbf{x}\in\mathscr{X}} \quad & \mathbf{c}^T\mathbf{x} + \sup \quad \sum_{k=1}^{K} p_k \; g_k(\mathbf{x}), \\ \text{subject to} \quad & \mathbf{p} \in \mathscr{P}, \end{aligned} \tag{B.20}$$

where $\mathscr{P} = \{\mathbf{p} : \mathbf{e}^T\mathbf{p} = 1,\ \mathbf{p} \geq 0,\ \sum_{k=1}^{K} \hat{p}_k \phi(\frac{p_k}{\hat{p}_k}) \leq \rho\}$. This is equivalent to minimizing $\mathbf{c}^T\mathbf{x} + \beta$, where $\beta \geq \mathbf{p}^T\mathbf{g}(\mathbf{x})$ for all $\mathbf{p} \in \mathscr{P}$. This robust constraint can be reformulated according to Theorem 1 in Ben-Tal et al. (2013). According to the theorem, $\mathbf{x}$ satisfies such an infinite dimensional constraint if and only if there exists $\eta, \lambda \in \mathbb{R}$ such that $(\mathbf{x}, \eta, \lambda)$ satisfy $\beta \geq \eta + \rho\lambda + \lambda \sum_{k=1}^{K} \hat{p}_k \phi^\star\left(\frac{g_k(\mathbf{x})-\eta}{\lambda}\right)$, $\lambda \geq 0$. Therefore, we can replace

(B.20) with:

$$\begin{aligned} \min_{\mathbf{x}\in\mathscr{X}} \quad & \mathbf{c}^T\mathbf{x} + \eta + \rho\lambda + \lambda\sum_{k=1}^{K}\hat{p}_k\phi^\star\left(\frac{g_k(\mathbf{x})-\eta}{\lambda}\right) \\ & \lambda \geq 0, \end{aligned} \tag{B.21}$$

where $0\phi^\star(b/0) = 0$ if $b \leq 0$ and $0\phi^\star(b/0) = \infty$ if $b > 0$. As before, we have an additional dual feasibility constraint to ensure that $\phi^\star$ is finite. We require that $g_k(\mathbf{x}) - \eta \leq \left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda$, for all $k = 1, ..., K$, for any ϕ-divergence for which the limit is finite.

As before, we require Assumption 2 to hold, so that $\phi^\star$ is monotone increasing, and so $\phi^\star(\min\ G(z_k, \mathbf{z}_k, \mathbf{Z}_k)) = \min\phi^\star(G(z_k, \mathbf{z}_k, \mathbf{Z}_k))$ for any function G. Thus, we can reformulate (B.21) as follows:

$$\begin{aligned} \min_{\mathbf{x}\in\mathscr{X},\eta,\lambda,\{z_k,\mathbf{z}_k,\mathbf{Z}_k,\mathbf{w}_k\}} \quad & \mathbf{c}^T\mathbf{x} + \eta + \rho\lambda + \lambda\sum_{k=1}^{r}\hat{p}_k\phi^\star\left(\frac{z_k + {\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k + \mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T) - \eta}{\lambda}\right) \\ & + \lambda\sum_{k=r+1}^{K}\hat{p}_k\phi^\star\left(\frac{\mathbf{m}_k^T\mathbf{w}_k - \eta}{\lambda}\right), \\ \text{subject to} \quad & \boldsymbol{\xi}^T\mathbf{Z}_k\boldsymbol{\xi} + \mathbf{z}_k^T\boldsymbol{\xi} + z_k \geq Q(\mathbf{x},\boldsymbol{\xi}),\ \forall\boldsymbol{\xi}\in\Omega_k,\ \forall k = 1, ..., r, \\ & z_k + {\mathbf{z}_k}^T\hat{\boldsymbol{\mu}}_k + \mathbf{Z}_k\circ(\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T) - \eta \leq \left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda, \\ & \forall k = 1, ..., r, \\ & \mathbf{m}_k^T\mathbf{w}_k - \eta \leq \left(\lim_{t\to\infty}\frac{\phi(t)}{t}\right)\lambda,\ \forall k = r+1, ..., K, \\ & \mathbf{W}\mathbf{w}_k = \mathbf{h} - \mathbf{T}\mathbf{x},\ \forall k = r+1, ..., K, \\ & \mathbf{w}_k \geq 0,\ \forall k = r+1, ..., K, \\ & \lambda \geq 0. \end{aligned} \tag{B.22}$$

Now we consider the infinite dimensional constraints $\boldsymbol{\xi}^T\mathbf{Z}_k\boldsymbol{\xi} + \mathbf{z}_k^T\boldsymbol{\xi} + z_k \geq Q(\mathbf{x},\boldsymbol{\xi}),\ \forall\boldsymbol{\xi} \in \Omega_k,\ \forall k = 1, ..., r$. As $Q(\mathbf{x},\boldsymbol{\xi}) = \min_{\mathbf{w}_k}$, the constraint is equivalent to the statement:

$$\forall\boldsymbol{\xi}\in\Omega_k, \quad \exists\mathbf{w}_k \text{ s.t. } \boldsymbol{\xi}^T\mathbf{Z}_k\boldsymbol{\xi} + \mathbf{z}_k^T\boldsymbol{\xi} + z_k \geq \boldsymbol{\xi}^T\mathbf{w}_k,\ \forall k = 1, ..., r,$$

where $\mathbf{w}_k$ must satisfy $\mathbf{W}\mathbf{w}_k = \mathbf{h} - \mathbf{T}\mathbf{x}$, $\mathbf{w}_k \geq 0$. This can be rewritten as:

$$\begin{aligned} \inf_{\boldsymbol{\xi}\in\Omega_k} \quad \max_{\mathbf{w}_k \geq 0} \quad & z_k + \mathbf{z}_k^T\boldsymbol{\xi} + \boldsymbol{\xi}^T\mathbf{Z}_k\boldsymbol{\xi} - \boldsymbol{\xi}^T\mathbf{w}_k, \\ \text{subject to} \quad & \mathbf{W}\mathbf{w}_k = \mathbf{h} - \mathbf{T}\mathbf{x} \qquad\qquad \geq 0. \end{aligned} \tag{B.23}$$

We focus on reformulating the term on the left hand side. The inner maximization problem is simply a linear program. We associate dual variable $\mathbf{v}_k$ and replace the inner problem with its dual (equivalent by strong duality):

$$\begin{aligned} \inf_{\boldsymbol{\xi}\in\Omega_k} \quad \min_{\mathbf{v}_k} \quad & z_k + \mathbf{z}_k^T\boldsymbol{\xi} + \boldsymbol{\xi}^T\mathbf{Z}_k\boldsymbol{\xi} + \mathbf{v}^T(-\mathbf{h} + \mathbf{T}\mathbf{x}), \\ \text{subject to} \quad & \mathbf{v}^T\mathbf{W} \leq \boldsymbol{\xi}^T. \end{aligned} \tag{B.24}$$

The constraint is equivalent to $\mathbf{W}^T\mathbf{v} \leq \boldsymbol{\xi}$. Let $\mathbf{U} = \mathbf{W}^T$. Without loss of generality, we may assume that $\mathbf{W}$ is full row rank. Otherwise, there is either a redundant constraint which we may remove or the system of equations is inconsistent and there is no solution. Thus, $\mathbf{U}$ is full column rank. Therefore $\mathbf{U}$ has a Moore-Penrose pseudoinverse which is a left inverse, such that $\mathbf{U}^\dagger\mathbf{U} = \mathbf{I}$. Now let $\mathbf{y} = \mathbf{W}^T\mathbf{v}$. We can rewrite (B.24) as:

$$\begin{aligned} \inf_{\boldsymbol{\xi}\in\Omega_k} \quad \min_{\mathbf{y}} \quad & z_k + \mathbf{z}_k^T\boldsymbol{\xi} + \boldsymbol{\xi}^T\mathbf{Z}_k\boldsymbol{\xi} + (-\mathbf{h} + \mathbf{T}\mathbf{x})^T\mathbf{U}^\dagger\mathbf{y}, \\ \text{subject to} \quad & \mathbf{y} \leq \boldsymbol{\xi}. \end{aligned} \tag{B.25}$$

For a given $\boldsymbol{\xi}$, the inner minimization problem is a linear program in $\mathbf{y}$ with no constraints other than an upper bound on $\mathbf{y}$. Therefore, if the coefficient of $\mathbf{y}$ is positive, then the objective function is unbounded below as $\mathbf{y}$ is not bounded below. Therefore, we must have the vector coefficient $(-\mathbf{h} + \mathbf{T}\mathbf{x})^T\mathbf{U}^\dagger \leq 0$. Under this condition, the minimum of the linear function will occur when $\mathbf{y}$ is largest, $\mathbf{y} = \boldsymbol{\xi}$. Thus, we can simply plug this into (B.25) to get:

$$\inf_{\boldsymbol{\xi}\in\Omega_k} \quad z_k + \mathbf{z}_k^T\boldsymbol{\xi} + \boldsymbol{\xi}^T\mathbf{Z}_k\boldsymbol{\xi} + (-\mathbf{h} + \mathbf{T}\mathbf{x})^T\mathbf{U}^\dagger\boldsymbol{\xi}. \tag{B.26}$$

Therefore, the constraint is equivalent to $\boldsymbol{\xi}^T\mathbf{Z}_k\boldsymbol{\xi} + \mathbf{z}_k^T\boldsymbol{\xi} + z_k + (-\mathbf{h} + \mathbf{T}\mathbf{x})^T\mathbf{U}^\dagger\boldsymbol{\xi} \geq 0$, $\forall\boldsymbol{\xi} \in \Omega_k,\ \forall k = 1, ..., r$. This is now an infinite dimensional constraint over the region Ω_k. Therefore, the total problem is as in (8.6). □

B.6 Proof of Theorem 12

Proof. All of the steps in formulating the dual problem are the same as in Section 8.1, so we get a similar formulation as in (B.22). However, in the infinite dimensional constraints $\mathbf{h}^T\mathbf{Z}_k\mathbf{h} + \mathbf{z}_k^T\mathbf{h} + z_k \geq Q(\mathbf{x}, \mathbf{h}),\ \forall \mathbf{h} \in \Omega_k,\ \forall k = 1, ..., r$, the domain is for the constraint vector $\mathbf{h}$. Since we have removed the randomness in the dual problem, we revert to lowercase and bold to denote the vector $\mathbf{h}$.

Consider (8.19), which is simply a linear program in $\mathbf{w}$. We associate dual vector $\mathbf{p}$, and consider the minimum of the Lagrangian:

$$\min_{\mathbf{w} \geq 0} \quad \boldsymbol{\xi}^T\mathbf{w} + \mathbf{p}^T(\mathbf{h} - \mathbf{T}\mathbf{x} - \mathbf{W}\mathbf{w}). \tag{B.27}$$

This will be bounded if and only if $\boldsymbol{\xi}^T - \mathbf{p}^T\mathbf{W} \geq 0$. In this case, the minimum will occur at $\mathbf{w} = 0$, and be equal to $\mathbf{p}^T(\mathbf{h} - \mathbf{T}\mathbf{x})$. This gives us the dual function. Assuming that (8.19) has a finite optimal value, then strong duality holds. We then maximize the dual function over all $\mathbf{p}$ to get that the second stage problem is equivalent to:

$$\begin{aligned} Q(\mathbf{x}, \mathbf{h}) = \qquad & \max_{\mathbf{p}_k} \quad (\mathbf{h} - \mathbf{T}\mathbf{x})^T\mathbf{p}_k, \\ & \text{subject to} \quad \mathbf{W}^T\mathbf{p}_k \leq \boldsymbol{\xi}. \end{aligned} \tag{B.28}$$

where we use the subscript k to denote which constraint we are considering. Now we consider the infinite dimensional constraints $\mathbf{h}^T\mathbf{Z}_k\mathbf{h} + \mathbf{z}_k^T\mathbf{h} + z_k \geq Q(\mathbf{x}, \mathbf{h}),\ \forall \mathbf{h} \in \Omega_k,\ \forall k = 1, ..., r$. As $Q(\mathbf{x}, \mathbf{h}) = \max_{\mathbf{p}_k}$, the constraint is equivalent to the statement:

$$\mathbf{h}^T\mathbf{Z}_k\mathbf{h} + \mathbf{z}_k^T\mathbf{h} + z_k \geq (\mathbf{h} - \mathbf{T}\mathbf{x})^T\mathbf{p}_k,\ \forall \mathbf{h} \in \Omega_k,\ \forall \mathbf{p}_k,\ \forall k = 1, ..., r,$$

where $\mathbf{p}_k$ must satisfy $\mathbf{W}^T\mathbf{p}_k \leq \boldsymbol{\xi}$. This can be rewritten as:

$$\begin{aligned} \inf_{\mathbf{h} \in \Omega_k} \quad \inf_{\mathbf{p}_k} \quad & z_k + \mathbf{z}_k^T\mathbf{h} + \mathbf{h}^T\mathbf{Z}_k\mathbf{h} - (\mathbf{h} - \mathbf{T}\mathbf{x})^T\mathbf{p}_k, \\ \text{subject to} \quad & \mathbf{W}^T\mathbf{p}_k \leq \boldsymbol{\xi} \qquad\qquad\qquad\qquad\qquad \geq 0. \end{aligned} \tag{B.29}$$

Let $\mathbf{v}_k = \mathbf{W}^T\mathbf{p}_k$. Since $\mathbf{W}$ has full row rank, $\mathbf{U} = \mathbf{W}^T$ has full column rank and thus

has a left pseudoinverse, $\mathbf{U}^\dagger\mathbf{U} = \mathbf{I}$. We can replace (B.29) with

$$\begin{aligned} \inf_{\mathbf{h}\in\Omega_k} \quad \inf_{\mathbf{v}_k} \quad & z_k + \mathbf{z}_k^T\mathbf{h} + \mathbf{h}^T\mathbf{Z}_k\mathbf{h} - (\mathbf{h}-\mathbf{T}\mathbf{x})^T\mathbf{U}^\dagger\mathbf{v}_k, \\ \text{subject to} \quad & \mathbf{v}_k \leq \boldsymbol{\xi} \qquad\qquad \geq 0. \end{aligned} \tag{B.30}$$

As the inner minimization problem is linear, and $\mathbf{v}_k$ is bounded above, it will be bounded below if and only if the vector coefficient $-(\mathbf{h}-\mathbf{T}\mathbf{x})^T\mathbf{U}^\dagger \leq 0$, $\forall \mathbf{h} \in \Omega_k$. This is equivalent to $\mathbf{W}^\dagger(\mathbf{h}-\mathbf{T}\mathbf{x}) \geq 0$, $\forall \mathbf{h} \in \Omega_k$. If there is a single $\mathbf{h} \in \Omega_k$ for which this inequality does not hold, then (B.30) cannot hold. If it holds, then the infimum of the inner problem will occur when $\mathbf{v}_k = \boldsymbol{\xi}$, for any $\mathbf{h}$. Hence, we arrive at:

$$\inf_{\mathbf{h}\in\Omega_k} \quad z_k + \mathbf{z}_k^T\mathbf{h} + \mathbf{h}^T\mathbf{Z}_k\mathbf{h} - (\mathbf{h}-\mathbf{T}\mathbf{x})^T\mathbf{U}^\dagger\boldsymbol{\xi} \qquad \geq 0. \tag{B.31}$$

Therefore, the constraints amount to two infinite dimensional constraints, one of which (B.31) requires nonnegativity of a quadratic function in $\mathbf{h}$ over a given region Ω_k, and the other requires nonnegativity of a set of linear functions in $\mathbf{h}$ over Ω_k. The number of such functions will be equal to the dimension of $\mathbf{w}$, denoted by p (equal to the number of rows in $\mathbf{W}^\dagger$). Let $\boldsymbol{\omega}_i$ be the ith row of $\mathbf{W}^\dagger$, made into a column vector. The p constraints can be represented as $\boldsymbol{\omega}_i^T\mathbf{h} - \boldsymbol{\omega}_i^T\mathbf{T}\mathbf{x} \geq 0$, $\forall \mathbf{h} \in \Omega_k$. This is equivalent to $\min_{\mathbf{h}\in\Omega_k} \boldsymbol{\omega}_i^T\mathbf{h} - \boldsymbol{\omega}_i^T\mathbf{T}\mathbf{x} \geq 0$. For an ellipsoidal fragmentation, this is a minimum of a linear function over an ellipsoid. We know the minimum will occur on the boundary. For a given ellipsoid Ω_k, let the boundary be described by the set $\{\mathbf{h} \ : \ (\mathbf{h}-\boldsymbol{\nu}_k)^T\boldsymbol{\Lambda}_k^{-1}(\mathbf{h}-\boldsymbol{\nu}_k) = \delta_k^2\}$. We consider the function $F(\mathbf{h}) = (\mathbf{h}-\boldsymbol{\nu}_k)^T\boldsymbol{\Lambda}_k^{-1}(\mathbf{h}-\boldsymbol{\nu}_k) - \delta_k^2$, whose gradient will be a normal vector to the ellipsoid, as it is a level set of the function. The minimum of $\boldsymbol{\omega}_i^T\mathbf{h} - \boldsymbol{\omega}_i^T\mathbf{T}\mathbf{x}$ will occur when the negative of the gradient of this linear function is perpendicular to the ellipsoid. Since the gradient of F is the normal vector, we must simply find the value of $\mathbf{h}$ for which these two vectors are multiples of one another. We calculate $\nabla F = -2\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\nu}_k + 2\boldsymbol{\Lambda}_k^{-1}\mathbf{h} = -\kappa\boldsymbol{\omega}_i$, where $\kappa \geq 0$. We solve to find $\mathbf{h} = -\kappa\boldsymbol{\Lambda}_k\boldsymbol{\omega}_i + \boldsymbol{\nu}_k$. We also know that $\mathbf{h}$ must satisfy $(\mathbf{h}-\boldsymbol{\nu}_k)^T\boldsymbol{\Lambda}_k^{-1}(\mathbf{h}-\boldsymbol{\nu}_k) = \delta_k^2$. We plug this in to arrive at $(\kappa\boldsymbol{\Lambda}_k\boldsymbol{\omega}_i)^T\boldsymbol{\Lambda}_k^{-1}(\kappa\boldsymbol{\Lambda}_k\boldsymbol{\omega}_i) = \delta_k^2$, which gives us that the multiple $\kappa = \frac{\delta_k}{\sqrt{\boldsymbol{\omega}_i^T\boldsymbol{\Lambda}_k\boldsymbol{\omega}_i}}$, and the minimum occurs at $\mathbf{h}^\star = -\frac{\delta_k}{\sqrt{\boldsymbol{\omega}_i^T\boldsymbol{\Lambda}_k\boldsymbol{\omega}_i}}\boldsymbol{\Lambda}_k\boldsymbol{\omega}_i + \boldsymbol{\nu}_k$. Thus, the minimum value of the linear function will be $\boldsymbol{\omega}_i^T\mathbf{h}^\star - \boldsymbol{\omega}_i^T\mathbf{T}\mathbf{x} = -\delta_k\sqrt{\boldsymbol{\omega}_i^T\boldsymbol{\Lambda}_k\boldsymbol{\omega}_i} + \boldsymbol{\omega}_i^T\boldsymbol{\nu}_k - \boldsymbol{\omega}_i^T\mathbf{T}\mathbf{x}$. We follow the same

procedures as before to represent the other constraints for an ellipsoidal fragmentation. Consequently we get the formulation in (8.20). Note that all of the constraints are either linear or LMIs. The number of new linear constraints grows linearly with p, the dimension of the second stage variable $\mathbf{w}$.

□

Appendix C

Acronyms

C.1 Acronyms

Table C.1: Acronyms

Acronym	Meaning
MDRO	Moments-based Distributionally Robust Optimization
SAA	Sample Average Approximation
PRO	Probabilistic Robust Optimization
RF	Robust Fragmentation
SDP	Semidefinite Program
LP	Linear Program
LMI	Linear Matrix Inequality
DRO	Distributionally Robust Optimization
RO	Robust Optimization
SCP	Self-Concordant Program
EM	Expectation-Maximization
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
CVaR	Conditional Value-at-Risk