Applications of Optimal Transport Theory to Process Optimization and Monitoring

by

Sanjula Kammammettu

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Process Control

Department of Chemical and Materials Engineering
University of Alberta

Abstract

Process optimization and process monitoring form two major cornerstones of the field of process systems engineering. This thesis tackles selected problems from these foci through the lens of optimal transport theory, a powerful mathematical methodology that is receiving renewed attention in recent years. The optimal transport problem seeks to transport probability mass from one probability distribution to another at the least total cost. This thesis uses this underlying concept in three main ways.

Firstly, the optimal transport distance is used as a measure of similarity between probability distributions. This thesis explores the use of entropy-regularized optimal transport to accomplish optimal reduction of large datasets that may be further used for scenario-based stochastic optimization. Entropy regularization for optimal transport proves to be advantageous in this case due to the availability of a numerical iterative solution scheme, alleviating the curse of dimensionality encountered in large-dimensional optimization problems. This work is further extended to generate optimal scenario trees for multistage stochastic programming problems. Results from case studies demonstrated that the proposed algorithms provide an efficient, iterative method to reduce the computational burden in scenario-based stochastic optimization, while also preserving the solution quality.

Secondly, the optimal transport distance is used to construct ambiguity sets used in distributionally robust optimization. This thesis explores the use of optimal transport between Gaussian mixtures for distributionally robust optimization, which seeks to retain desirable features of both stochastic and robust optimization frameworks. In this work, the optimization problem is considered fraught with distributional ambiguity on multimodal uncertainty that is modeled as a Gaussian mixture. An optimal transport variant for Gaussian mixtures is further used to construct an ambiguity set of distributions around this reference model, and a tractable formulation is presented. The superior performance of this proposed formulation is contrasted with the established Wasserstein method on an illustrative study, as well as on a portfolio optimization problem. The thesis then uses the proposed formulation to tackle chance-constrained optimization in a distributionally robust setting, wherein the worst-case expected constraint violation is restricted to a user-defined limit. In a similar vein, this formulation is shown to outperform the conventional Wasserstein method.

Finally, optimal transport distance is used as a measure of similarity in a process monitoring framework. This thesis presents the applicability of the optimal transport distance as a metric for change-point and fault detection in multivariate processes and compares its performance with that of conventional fault detection metrics. The final component of this thesis tackles the fault detection problem through the lens of distributional ambiguity. In this work, distributional ambiguity is considered in the context of a multimodal process, and the worst-case performance of a fault detection system is evaluated on the basis of two performance metrics - false alarm rate, and fault detection rate. The evolution of worst-case performance metrics is tracked for varying levels of ambiguity, using the distributionally robust optimization formulation proposed in earlier chapters. The thesis concludes with a summary of the work conducted, the knowledge gaps addressed, and some future directions.

Preface

The work published in this thesis was conducted under the supervision of Dr. Zukui Li. The computational software platforms used for this work comprises GAMS, MATLAB and Python. The publication/submission details of the chapters are outlined below.

Chapter 1 of this thesis provides a background and motivation for the problems addressed in this thesis, in addition to an overview of the underlying concepts that connect the works undertaken.

Chapter 2 of this thesis has been published as "Scenario reduction and scenario tree generation for stochastic programming using Sinkhorn distance" in *Computers & Chemical Engineering*, 170 (2023): 108122 by Sanjula Kammammettu, and Zukui Li. The financial support for this work was provided by the Natural Sciences and Engineering Resource Council (NSERC) of Canada. The CRediT authorship contribution statement for this publication reads as follows: *Sanjula Kammammettu*: Methodology, Software, Writing - original draft; *Zukui Li*: Conceptualization, Supervision, Writing - review & editing, Funding Acquisition.

Chapter 3 of this thesis has been published as "Distributionally robust optimization using optimal transport for Gaussian mixture models" in *Optimization & Engineering* (2023): 1-26 by Sanjula Kammammettu, Shu-Bo Yang, and Zukui Li. The financial support for this work was provided by the Natural Sciences and Engineering Resource

Council (NSERC) of Canada. The authorship contribution statement for this publication reads as follows: Sanjula Kammammettu: Writing – review & editing, Writing – original draft, Software, Methodology, Data curation. Shu-Bo Yang: Writing – review & editing, Methodology. Zukui Li: Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. This work has been published as a part of this thesis only.

Chapter 4 of this thesis has been published as "Distributionally Robust Chance-Constrained Optimization with Gaussian Mixture Ambiguity Set" in *Computers & Chemical Engineering* (2024): 108703 by Sanjula Kammammettu, Shu-Bo Yang, and Zukui Li. The financial support for this work was provided by the Natural Sciences and Engineering Resource Council (NSERC) of Canada. The CRediT authorship contribution statement for this publication reads as follows: Sanjula Kammammettu: Writing – review & editing, Writing – original draft, Software, Methodology, Data curation. Shu-Bo Yang: Writing – review & editing, Methodology. Zukui Li: Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. This work has been published as a part of this thesis only.

Chapter 5 of this thesis has been published as "Change point and fault detection using Kantorovich Distance" in the *Journal of Process Control*, 80 (2019): 41-59 by Sanjula Kammammettu and Zukui Li. The financial support for this work was provided by the Natural Sciences and Engineering Resource Council (NSERC) of Canada Discovery Grant Program. The authorship contribution statement for this publication reads as follows: *Sanjula Kammammettu*: Methodology, Software, Writing - original draft; *Zukui Li*: Conceptualization, Supervision, Writing - review & editing, Funding Acquisition.

Chapter 6 of this thesis in a manuscript under preparation titled "Performance evaluation of a multimodal process fault detection system under distributional ambiguity" by Sanjula Kammammettu and Zukui Li. The financial support for this work was provided by the Natural Sciences and Engineering Resource Council (NSERC) of Canada Discovery Grant Program. The authorship contribution statement for this publication reads as follows: *Sanjula Kammammettu*: Methodology, Software, Writing - original draft; *Zukui Li*: Conceptualization, Supervision, Writing - review & editing, Funding Acquisition.

“I say not that it is, but that it seems to be; as it now seems to me to seem to be.”

- Hubert N. Alyea

To my family - who imbued me with the freedom, and confidence to do things my way

Acknowledgements

At the outset, I would like to express my gratitude to my supervisor, Dr. Zukui Li. Dr. Li - it has been a privilege to work with you, and learn from you. I thank you for your steady support and guidance during my academic journey, and for allowing me the space for growth, exploration and development, not just in research and academics, but also on a professional front. I would also like to thank my research group members - Shu-Bo, and Zhongyu - for their feedback and suggestions during my program.

Being a part of the CME department at UAlberta was an enriching experience, and I am fortunate to have learnt and developed my co-curricular and extracurricular skills, as a result of being amidst the professors and staff here. In this regard, I would like to extend my gratitude to Dr. Vinay Prasad, and the department administration for their encouragement, support, and above all, a willingness to engage with me on my duties as the CMEGSA President. I would also like to thank the CMEGSA Team 2023-24 for their work and company.

I would also like to thank the various course instructors - Dr. Samir Mushrif, Dr. Mohd Adnan Khan, Dr. Fereshteh Sattari, and Dr. Zukui Li - who have afforded me the opportunity to hone my teaching and communication skills. I would like to thank Dr. Muddu Madakyaru, my first guide in the world of process systems engineering, and a valuable mentor. I would also like to thank V. Vijaykiran sir for being my first inspiration to pursue teaching, and exemplifying the power of engagement in the classroom.

I would like to thank all those who have been my companions in this graduate school experience. I have had the good fortune of great friends, and this journey has been made richer by their presence and impact in my life. To Karthik, Bhubesh, and Sagar - thank you for all the conversations, the many laughs, and much-needed drives; DICE Level 3 was home largely due to your friendship, and it has been a great comfort to have you just a few desks away. To Hareem, and Sanober - you bridged the miles between us with your constant support and steadfastness; thank you for celebrating all my wins, and sharing in this experience. To Khyati, Hemanth, Shweta, Nilesh, and Ananthan - thank you for always being one car-ride, one flight, one phone-call away, and for pulling me out of those not-insignificant slumps. To those who have accompanied me on countless 'mental health walks'; to those who shared their company and time at tea-time; and to all those at DICE who always kept my spirits up - thank you.

And finally, I would like to thank my parents, and my sister for being my unwavering support system and cheerleaders through this journey, and whose belief in my abilities, and confidence in my decisions has been a constant source of strength for me.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Decision-making is an inherent part of numerous process operations and industries, such as manufacturing, mining, logistics, finance, economics, risk management and disaster response. In recent decades, the task of making optimal decisions has become increasingly automated, and therefore, developing a mathematical model that describes the true system well is a key task. Since the decisions prescribed as a result of simulating such optimization models for required operating conditions must finally be applied to the real system, it is imperative that these decisions be feasible in the face of real-world uncertainties (Edgar *et al.* 2001; Sahinidis 2004; Ning and You 2019; Keith and Ahner 2021). The development of probability theory in the 20th century has provided researchers with a rigorous mathematical framework using which mathematical definitions of uncertainty may be incorporated into optimization models. With continual progress in the development of new mathematical techniques, bolstered by improvement in computational power, optimal decision-making under uncertainty is now a multifaceted, multidisciplinary field that is fast-evolving. In this context, a host of mathematical research problems, especially in science, engineering, and economics, have been created that address a number of end-use cases and applications.

Process monitoring forms another major component of process systems engineering. The automation of industrial processes has brought about a number of changes in the protocols and operation of daily activities; significant among these is the task of monitoring processes to detect uninitiated changes in operation, and abnormal events (also termed faults). In the past, human operators formed the main supervisory control layer in industry. However, with time, industries have grown larger in scale and complexity, and their data is multidimensional and often correlated. Therefore, industries have begun to move away from relying on a manual process monitoring framework to ensure normal operation. To this end, abnormal event management (AEM) is a component of the process control hierarchy, and involves the automated tasks of detecting faults, diagnosing their origins in the system, and providing recommendations for remedial measures (Isermann 1984; Wise and Gallagher 1996; Qin 2012; Ge *et al.* 2013; Severson *et al.* 2016). The primary step in monitoring a process for normal operation is to ensure that the system operates only at conditions prescribed by the operator. This requires that any changes in the process, or any abnormal events that move the process away from its normal operating conditions be detected in a timely fashion. In literature, this is referred to as the "change point detection" or "fault detection" problem. The performance of a fault detection system is primarily assessed through its fault detectability, and the false alarm rate. Fundamentally, a system designed for high fault detectability is also susceptible to more false alarms; therefore, an optimal design of a fault detection system is a topic of interest in the process monitoring community.

This thesis addresses problems in the broad areas of mathematical optimization methods under uncertainty, and process monitoring. The chapters in this thesis are connected by the common thread of the underlying mathematical concept, namely optimal transport theory. The objectives of this thesis may be broadly classified into three categories,

1. To propose improvements for enhanced computational performance of a traditional method for optimization under uncertainty, namely stochastic programming

2. To propose mathematical formulations for a newer approach to optimization under uncertainty, called distributionally robust optimization, for uncertainty that exhibits multimodal characteristics

3. To address aspects of the process monitoring (fault detection) problem via optimal transport theory

The rest of this chapter provides an introduction to the different frameworks of optimization under uncertainty explored in this thesis, and where they fit into its considerations. This chapter further expands on optimal transport theory (the conceptual link) that forms the basis of the methods and formulations proposed in this thesis. It may be noted that a rigorous mathematical treatment of the methods and concepts is not presented in this chapter; however, all pertinent mathematical formulations have been derived and discussed in detail in each following chapter. It may be noted that some background and theory is repeated for completeness in each chapter that has been published independently.

1.2 Background theory

1.2.1 Optimization under uncertainty

Real world process optimization is riddled with uncertainty arising from a number of factors such as pricing, supply and demand targets, modeling approximations, and unmodeled disturbances. To this end, while deterministic formulations of the problem may offer "optimal" solutions, these decisions often fail in practice due to their disregard of uncertainty in the modelling and solution process. Optimization under uncertainty has been a significantly studied field since the 20th century wherein researchers noticed that practical applications of optimization were almost always bottlenecked by uncertainty (Sahinidis 2004). A large amount of research has gone into the fundamental concepts of mathematical modelling of the problem accounting for uncertainty, addressing computational intractability arising due to this, and finding solution procedures for the same. This section introduces an overview of the traditional approaches to optimization under uncertainty, namely stochastic programming, chance-constrained programming, and robust optimization that are distinguished by their consideration of uncertainty and its characterization in the model. It further introduces the distinction between uncertainty and "ambiguity" to give an overview of the relatively newer paradigm in optimization under uncertainty that offers solutions with "distributional robustness". Throughout this section, uncertainty is denoted by ξ, and the probability distribution governing it is denoted as $\mathbb{P}(\xi)$. While the general mathematical formulations are given where relevant, a detailed derivation of the approaches studied in this thesis may be found in their associated chapters.

1.2.1.1 Two-stage stochastic programming

Stochastic programming (SP) models are mathematical formulations of optimization models accounting for uncertainty (that is, risk) by leveraging information about the probability distribution governing the uncertainty in the problem (Dantzig 1955; Beale 1955). Under an objective cost minimization framework, SP models find optimal decisions under uncertainty by minimizing the "expected" cost under the distribution $\mathbb{P}(\xi)$.

SP models are commonly used in literature in the context of stage-wise scenario-based programming, wherein a finite set of scenarios approximates the support of the underlying distribution $\mathbb{P}(\xi)$. The scenario-based approach offers attractive computational advantages owing to its linear programming structure for a piece-wise linear function $f(x,\xi)$. The two-stage scenario-based SP model is a classic case of this type. Here, first-stage decisions are made independent of uncertainty, while second-stage decisions are taken as a recourse to the realization of uncertainty between these stages. To this end, first-stage decisions are static, while recourse decisions are adaptive to the realized uncertainty. The general form of the two-stage stochastic programming problem may be given as (Shapiro and Philpott 2007),

$$\min_{x\in X} \;\; c^{\mathrm{T}}x + \mathbb{E}_{\mathbb{P}(\xi)}\big[Q(x,\xi)\big] \tag{1.1}$$

where $Q(x,\xi) \coloneqq \min\limits_{y} \;\big\{q^{\mathrm{T}}y \;\mid\; Tx + Wy \le h\big\}$, and $\xi = (q,T,W,h)$. The $K-$scenario-based formulation of the stage-wise SP model in 1.1 may be given as,

$$\min_{x\in X, y_k|_{k=1}^{K}} \;\; c^{\mathrm{T}}x + \sum_{k=1}^{K} q_k^{\mathrm{T}} y_k \tag{1.2a}$$

$$\text{s.t.} \;\; T_k x + W_k y_k \le h_k, \quad \forall k = 1,\dots,K \tag{1.2b}$$

A natural extension to the two-stage SP model is the multi-stage SP model wherein uncertainty is gradually realized successively over time.

Scenario-based stochastic programming, through its definition, is risk-neutral in its formulation. A number of works suggest that the distribution $\mathbb{P}(\xi)$ be approximated by an empirical uniformly-weighted discrete distribution. This technique is popularly referred to as "sample average approximation" (SAA) in literature (Ahmed and Shapiro 2002; Luedtke and Ahmed 2008). Consequently, the performance of the SP model is significantly affected by this fitted empirical distribution. Compiling a scenario set for SP is therefore a significant part of the optimization process, and a considerable amount of research has gone into the same (Dantzig and Infanger 1991; Dempster and Thompson 1999; Høyland and Wallace 2001; Pflug 2001; Høyland *et al.* 2003; Heitsch and Römisch 2009b). While it is agreed upon that a larger set of scenarios provides more information on the true nature of $\mathbb{P}(\xi)$, leading to better solutions, incorporating too many scenarios into the SP model renders it too large and possibly intractable. Therefore, it is imperative that a scenario set be optimally sized with a good amount of information on $\mathbb{P}(\xi)$ (Heitsch and Römisch 2003; Dupačová *et al.* 2003). Chapter 2 of this thesis deals with the reduction of a large (super)set of scenarios to an optimal subset using an entropy-regularized variant of optimal transport theory, as described in further sections. In this chapter, the idea of scenario reduction is also extended to multistage scenario tree generation from a large number of time-varying uncertainty profiles.

1.2.1.2 Chance-constrained programming

Scenario-based stochastic programming deals with the minimization of the expected objective cost function value subject to a set of discrete realizations (or scenarios) of uncertainty (ξ), wherein all constraints in the model must be satisfied for all realizations in this scenario set.. This introduces a possibility of model infeasibility. To this end, chance-constrained programming (CCP) is a technique of optimization under uncertainty developed to tackle this issue of constraint satisfaction (Charnes and Cooper 1959). Specifically, in CCP, a "probabilistic" constraint is introduced wherein a certain small level of constraint violation is permissible, thus broadening the feasible space for the decision variables. That is, for the general optimization problem under uncertainty of the form,

$$\min_{x \in X} \quad f(x) \tag{1.3a}$$

$$\text{s.t.} \quad g_i(x,\xi) \leq 0 \quad \forall i = 1,\ldots,m, \quad \forall \xi \tag{1.3b}$$

the general form of the $m-$joint chance-constrained programming problem (Miller and Wagner 1965) may be given as,

$$\min_{x \in X} \quad f(x) \tag{1.4a}$$

$$\text{s.t.} \quad \Pr\Big\{ g_i(x,\xi) \leq 0 \quad \forall i = 1,\ldots,m \Big\} \geq 1 - \delta \tag{1.4b}$$

It may be noted that while chance-constrained programming introduces a certain relaxation in the constraints of the problem that aid in finding a feasible solution, the probabilistic constraint involved with CCP also introduces computational intractability in its general form. In practice, this probabilistic constraint is often reformulated to an indicator function-based form of the expectation under $\mathbb{P}(\xi)$, and further approximated by a convex, conservative approximation to the same (Ben-Tal and Nemirovski 2000; Rockafellar, Uryasev, *et al.* 2000; Nemirovski and Shapiro 2007) or using sample average approximation (Luedtke and Ahmed 2008). Furthermore, as in the case of stage-wise stochastic programming, the probabilistic constraint in CCP requires the knowledge of the distribution $\mathbb{P}(\xi)$. Chapter 4 contains a mathematical discussion on CCP, as well as its extension to the distributionally robust framework to hedge against the lack of exact information on $\mathbb{P}(\xi)$.

1.2.1.3 Robust optimization

While stochastic programming techniques offer formulations that leverage distributional information for optimization under uncertainty, these methods are also limited by the quality of data, and consequently, the empirical distribution fitted. In many industrial applications, particularly in the case of design, as well as safety features, this dependence on the quality of discrete scenarios is limiting, and a safer, conservative solution is preferable. To this end, robust optimization (RO) techniques optimize for the worst-case realization of the problem across a feasible uncertainty set, and require no knowledge on the distribution $\mathbb{P}(\xi)$. It is noteworthy to mention that the robustness of the solution may refer to its feasibility towards different realizations of uncertainty, or even objective value or optimality guarantees. However, RO requires the careful design of an uncertainty set in order to avoid overly conservative decisions that come at an unnecessarily higher cost to the end user. Furthermore, disregarding information on $\mathbb{P}(\xi)$ entirely may cause the loss of relevant and useful features for optimization.

In order to ensure feasibility of the solution over all realizations of the uncertain parameter in the model, in practice, a limited "uncertainty set" of scenarios is constructed around nominal information available on the parameter. The pioneering work in RO was undertaken by Soyster (1973). A major component of RO research has been undertaken on solving the worst-case optimization problem. Specifically, this refers to the class of methods that minimize the objective based on the worst-case feasible solutions that are realized from the uncertainty set constructed. The general form of the RO problem may be given as,

$$\min_{x \in X} \max_{\xi \in \Xi} \; f(x, \xi) \tag{1.5a}$$

$$\text{s.t.} \;\; g_i(x, \xi) \leq 0, \quad \forall i = 1, ..., m, \quad \forall \xi \in \Xi \tag{1.5b}$$

Here, Ξ denotes the uncertainty set. It is worth noting that Model 1.5 finds optimal one-time decisions with no recourse. Similar to stage-wise stochastic programming, RO can also be tackled in an "adjustable" form wherein some decision variables may be defined as "wait-and-see" decisions that may be arrived at once the value of the uncertain parameter is realized at a future time (Ben-Tal *et al.* 2004; Boni and Ben-Tal 2008; Yanıkoğlu *et al.* 2019). A key factor to be addressed in RO is the tradeoff between robustness and conservatism of the solution that is to be tuned for best performance (Bertsimas and Sim 2004).

This thesis does not explicitly focus on robust optimization methods. However, it leverages the worst-case approach that is typical of RO, to tackle a "distributionally robust" version of the optimization problem under uncertainty, as detailed in the next section.

1.2.1.4 Uncertainty vs ambiguity - extended approaches to optimization under uncertainty

Optimization under uncertainty has been a topic of interest in a myriad of settings ranging from process optimization and design, to profit maximization and portfolio design in economic studies. The latter field of research has provided a number of conceptual insights into the incorporation of uncertainty in system modeling and optimization. Specifically, the concepts of uncertainty, risk, and ambiguity have been refined by various researchers over the years, and different techniques to address the same have been continually studied. In this context, Keynes (1921) and Knight (1921) separately made the primary characterization of uncertainty and risk; however, in their works, the term "uncertainty" refers to all models containing parameters with unknown probability distributions. An analysis, and comparison of their philosophies and treatment of uncertainty in a system is presented in Packard *et al.* (2021). While these seminal works lay the ground for the philosophy of uncertainty, Arrow (1951) made a clear distinction between the types of uncertainty a user may encounter in the model, namely, the uncertainty in the *hypothesis*, and the uncertainty in the *observations*, given the hypothesis. The uncertainty in the hypothesis is more commonly referred to now as "ambiguity", and refers to the uncertainty on the probability distribution itself, while the uncertainty in observations is expressed by that probability distribution. It follows that if one is "certain" about the probability distribution (that is, the hypothesis is a certainty), ambiguity may be disregarded, and the uncertainty in realizations/observation may be addressed by the aforementioned traditional approaches.

Accounting for ambiguity in the distribution $\mathbb{P}(\xi)$ provides a new outlook to optimization under uncertainty, which may be summarized under two main points. Firstly, accounting for ambiguity in $\mathbb{P}(\xi)$, particularly in the face of limited amounts of data from which statistical properties may be gleaned, ensures a degree of practicality into the optimization model. Secondly, accounting for ambiguity in $\mathbb{P}(\xi)$ has allowed for a new paradigm in optimization methods under uncertainty, called "distributionally robust optimization" (DRO) that is able to retain the positive qualities of the traditional stochastic programming (SP) and robust optimization (RO) approaches, while simultaneously alleviating their drawbacks. Indeed, DRO is also referred to as "ambiguous stochastic optimization" in literature, and may be considered an intermediate approach to SP and RO. The first instances of DRO in literature were published in the context of a newsvendor problem by Karlin *et al.* (1958) and on generic linear programming problems by Žáčková (1966). The DRO problem aims to minimize the supremum, or the worst-case realization, of the expected cost due to uncertainty in the problem, over a defined "ambiguity set" containing a number of possible candidate distributions that the uncertain parameter may follow. DRO retains elements of SP by incorporating the use of $\mathbb{P}(\xi)$ into the problem. However, unlike traditional SP, DRO does away with the certainty on $\mathbb{P}(\xi)$, and introduces a degree of belief on $\mathbb{P}(\xi)$ through the ambiguity set that is usually constructed using available knowledge on ξ. On the other hand, DRO retains the worst-case approach of RO by hedging, not against a single $\mathbb{P}(\xi)$, but against that distribution from amongst all the candidate distributions in the ambiguity set that realizes the worst-case performance of the model. The general formulation of the DRO problem may be given as,

$$\min_{x \in X} \max_{\mathbb{P}(\xi) \in \mathcal{P}} \mathbb{E}_{\mathbb{P}(\xi)}\big[f(x, \xi)\big] \tag{1.6}$$

where $\mathcal{P}$ denotes the ambiguity set containing all hypothesized models or probability distributions on ξ. A number of review papers are available on DRO, that classify and summarize work done by researchers over the years on the different formulations, algorithms, and ambiguity set construction procedures (Bayraksan and Love 2015; Postek *et al.* 2016; Shapiro 2021; Lin *et al.* 2022). Specifically, the construction of the ambiguity set for DRO is a significant consideration for solution quality. A number of works in literature have focused on methods for ambiguity set construction (Ghaoui *et al.* 2003; Popescu 2005; Pflug and Wozabal 2007; Delage and Ye 2010; Hu and Hong 2013; Hanasusanto and Kuhn 2013; Bayraksan and Love 2015; Ning and You 2018; Li 2018; Shang and You 2018; Esfahani and Kuhn 2018; Chen 2018; Chen *et al.* 2022).

All of these methods construct the ambiguity set around a "nominal" distribution that is empirically determined from available data on ξ. Chapters 3 and and 4 of this thesis tackle the DRO problem for multimodal uncertainty. Chapter 3 tackles the DRO problem to find the optimal solution to the worst-case expectation maximization problem, which is the inner maximization problem applied to the expected value of the objective function under uncertainty. Chapter 4 tackles the distributionally robust chance-constrained problem (DRCCP) using the results of Chapter 3. Chapter 6 extends the methodology proposed in Chapter 3 to process monitoring applications in the context of optimal fault detection system design.

1.2.2 Optimal transport theory

The basic foundations of optimal transport theory were laid against the backdrop of the French Revolution of the 18^{th} century, when Gaspard Monge, the inventor of descriptive geometry and a mathematician working with the French military body was tasked with the following problem statement, "what is the most cost-effective method to shape a pile of dirt into a desired form?". This problem underwent a number of rebirths from the 18^{th} century to the present day, giving rise to different formulations, computational solution strategies, and a host of new applications as other fields of science and technology progressed. The seemingly logistical nature of the original problem statement is now far-reaching; in contemporary literature, optimal transport forms a strong basis for a host of machine learning and artificial intelligence applications including image and computer vision, and generative networks.

1.2.2.1 Mathematical formulations of the optimal transport problem

From its inception through its evolution in mathematical literature, the aim of the optimal transport problem has been to obtain the least-cost method of transporting probability mass from a source distribution to a destination distribution. For the source and destination distributions $\mathbb{P}(x)$ and $\mathbb{Q}(y)$ on supports X and Y, respectively, Monge's pioneering formulation (Monge 1781) of the optimal transport (OT) problem may be defined as,

$$M(\mathbb{P},\mathbb{Q}) \coloneqq \inf_{f\in\mathcal{M}} \left\{ \int_X c\big(x, f(x)\big) \ \ d\mathbb{P}(X) \right\} \tag{1.7}$$

wherein $M(\mathbb{P},\mathbb{Q})$ is a continuous transport map between $\mathbb{P}$ and $\mathbb{Q}$. $c(x, f(x))$ denotes the cost incurred during the transport of probability mass from $\mathbb{P}$ to $\mathbb{Q}$ computed between the elements of their support sets; in his original work, Monge computed the cost using the L_1 norm. $\mathcal{M}$ refers to the set of measure-preserving mappings from X to Y for any Borel subset A of Y as,

$$\mathcal{M} = \left\{ f : X \to Y \middle| \int_{f^{-1}(A)} d\mathbb{P}(x) = \int_A d\mathbb{Q}(y) \right\} \tag{1.8}$$

Solving the OT problem using Monge's formulation in Model 1.7 poses a number of practical difficulties owing to its "pushforward" definition of a transport map which induces nonlinearity. Furthermore, in this definition of the OT problem, it is worth noting that probability mass is mapped from one distribution to another (Kantorovich 1942); this means that it may not be split during transport. This inherent assumption in Model 1.7 poses difficulties regarding the existence of such transport maps when the problem is treated in a discrete capacity. To solve this issue, Leonard Kantorovich posed the relaxed version of the OT problem wherein probability mass is "transferred" from one distribution to another, allowing for mass from the source to be split across multiple supporting locations in the destination. The Kantorovich optimal transport problem may be defined as,

$$K(\mathbb{P},\mathbb{Q}) \coloneqq \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \left\{ \int_{X\times Y} c\big(x, y\big) \ \ d\gamma(x, y) \right\} \tag{1.9}$$

wherein $K(\mathbb{P},\mathbb{Q})$ is referred to as the optimal transport "plan" of probability mass from $\mathbb{P}$ to $\mathbb{Q}$. $\Gamma(\mathbb{P},\mathbb{Q})$ refers to the set of all joint distributions whose marginal distributions are $\mathbb{P}$ and $\mathbb{Q}$; that is, for all measurable sets $A \subseteq X$ and $B \subseteq Y$,

$$\Gamma = \left\{ \gamma : \gamma(A, Y) = \mathbb{P}(A), \gamma(X, B) = \mathbb{Q}(B) \right\} \tag{1.10}$$

When $\mathbb{P}$ and $\mathbb{Q}$ are absolutely continuous with respect to the Lebesgue measure, it has been proven that a unique solution to γ may be obtained (Brenier 1991; Villani 2021). It may be shown that when transport maps do exist between probability distributions $\mathbb{P}$ and $\mathbb{Q}$, the optimal transport "distance", which is the total cost of transporting probability mass from $\mathbb{P}$ to $\mathbb{Q}$, obtained from the Kantorovich formulation serves as a lower bound to that of the Monge OT problem. It may further be noted that the Kantorovich problem in Model 1.9 is a convex optimization problem in that, (1) the constraints involved in the problem are convex, and (2) the cost function, which is usually norm-based or distance-based, is also convex. The Kantorovich OT problem also naturally admits a linear programming (LP) formulation in the discrete case; insofar as the application of optimal transport theory to the problem statements addressed in this thesis, the discrete OT problem forms the starting point. For the discrete measures $\zeta = \sum_{i=1}^{m} a_i \delta_{x_i}$ and $\omega = \sum_{j=1}^{n} b_j \delta_{y_j}$ wherein the support elements are modeled as Dirac point masses δ_{x_i} and δ_{y_j} with weights a_i and b_j respectively, the discrete formulation of the Kantorovich OT problem is given as,

$$\min_{\pi} \quad \sum_{i=1}^{m} \sum_{j=1}^{n} c_{i,j} \pi_{i,j} \tag{1.11a}$$

$$\text{s.t.} \quad \sum_{j=1}^{n} \pi_{i,j} = a_i, \quad \forall i = 1, ..., m \tag{1.11b}$$

$$\sum_{i=1}^{m} \pi_{i,j} = b_j, \quad \forall j = 1, ..., n \tag{1.11c}$$

$$\pi_{i,j} \geq 0, \quad \forall i = 1, ..., m, \quad j = 1, ..., n \tag{1.11d}$$

where π refers to the optimal transport plan. Inherent in this formulation is the assumption that the weights a_i and b_j of the Dirac point supports for ζ and ω, respectively, are nonnegative, and have unitary sum.

1.2.2.2 Optimal transport distances

While the optimal transport (OT) problem was founded owing to the need to find a least cost "method", which may be a map or a plan, to move probability masses from one measure to another, a significant number of advances have been made in literature whose focus is on the subsequent cost due to this method, rather than this transport method itself. The objective function value of the OT problem, which denotes the total cost of transporting probability mass from the source $\mathbb{P}(x)$ to the destination $\mathbb{Q}(y)$ as a result of following the transport plan $K^*(\mathbb{P},\mathbb{Q})$ is referred to as the optimal transport (OT) "distance" between $\mathbb{P}$ and $\mathbb{Q}$; simply put, the OT distance is an aggregate cost incurred by transporting probability mass between $\mathbb{P}$ and $\mathbb{Q}$.

The OT distance so defined from the optimal transport problem finds notable application as a metric of similarity between probability distributions. The p–Wasserstein distance is such a metric defined as the p^{th} root of the optimal objective cost (that is, the OT distance) of the Kantorovich OT problem between measures ζ and ω which have bounded p^{th} moments. The p–Wasserstein distance, denoted as $W_p(\zeta,\omega)$, may be calculated as,

$$W_p(\zeta,\omega) = \min_{\pi\in\Pi(\zeta,\omega)} \left(\int_{X\times Y} |x-y|^p \quad d\pi(x,y) \right)^{\frac{1}{p}} \tag{1.12}$$

It has been proven that $W_p(\zeta,\omega)$ is a metric using the triangle inequality. The use of W_p as a metric has found considerable applications in the fields of computer vision and machine learning, as a method of comparison of histograms. It has also be used for image analysis and signal processing along similar lines. This thesis explores the use of the OT/W_p distance, or more specifically its extensions in different contexts. Chapter 2 treats the support of the destination distribution $\mathbb{Q}(y)$ as a subset of that of the source $\mathbb{P}(x)$. Chapters 3 and 4 construct a number of distributions around the source $\mathbb{P}(x)$ constrained by a defined OT distance to propose extended OT-distance-based distributionally robust frameworks for optimization. Chapter 5 explores the use of the OT distance as a metric for similarity for process change point and fault detection. Chapter 6 presents a study of the worst-case performance of a fault detection system using the frameworks proposed in Chapters 3 and 4.

1.2.2.3 Extensions to the conventional OT problem

Since its inception in the 18$^{\text{th}}$ century, the field of optimal transport (OT) theory has undergone a number of changes and rebirths under varying formulations and implementation strategies. In recent years, interest in the use of OT has experienced yet another regeneration owing to the development and easy availability of new solvers that can handle large high-dimensional datasets. This interest is further reinforced due to the current availability of large amounts of data that may be leveraged for decision-making. To this end, a popular extension to the OT problem that is often used in a myriad of fields from computer vision to machine learning is the entropy-regularized variant, also called Sinkhorn optimal transport. Chapter 2 of this thesis uses Sinkhorn optimal transport to address the optimal scenario reduction, and scenario tree generation problem for stochastic programming. While many other variants, such as multi-marginal, unbalanced, and semi-discrete versions of OT exist, in this section, Chapters 3 and 4 focus on modeling multimodal uncertainty in optimization problems, and to this end, the use of a Gaussian mixture model-based variant of OT is used in order to design components of the distributionally robust optimization frameworks in Chapters 3 and 4.

1.2.2.3.1 Entropy-regularized (Sinkhorn) optimal transport

The increase in the availability of computational power in recent decades has led researchers to explore numerical approximations to the Kantorovich optimal transport (OT) problem. One such approximation was made through an "entropic regularization penalty" added to the objective function in the OT problem. This regularization leads to a reformulated OT problem that may be easily solved iteratively by means of the Sinkhorn-Knopp algorithm; therefore, entropy-regularized OT is also referred to as the Sinkhorn optimal transport problem. The theory and numerical scheme of Sinkhorn optimal transport has been well-explained in the introduction of Chapter 2. This section provides an overview of the same.

The idea of introducing entropy regularization into the original OT problem was originally used in transportation theory by Wilson (1969) and Erlander (1980). In these works, it was observed that transport problems tend to have more diffused solutions in practice than those given by OT theory which by nature offers more sparse couplings; to this end, an entropy-regularized version of OT was developed seeking these diffused couplings. The discrete entropy of the optimal transport plan may be given as,

$$\mathbb{H}(\pi) := -\sum_{i=1}^{m}\sum_{j=1}^{n}\pi_{i,j}\big[\log(\pi_{i,j}) - 1\big] \tag{1.13}$$

When (negative) discrete entropy $-\mathbb{H}(\pi)$ is added to the objective function of the optimal transport problem in Model 1.11, it serves as a regularization function, and thus, the entropy-regularized variant of the OT problem may be given as,

$$\begin{aligned}
\min_{\pi} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} c_{i,j}\pi_{i,j} + \gamma\pi_{i,j}\big[\log(\pi_{i,j})\big] && \text{(1.14a)}\\
\text{s.t.} \quad & \sum_{j=1}^{n}\pi_{i,j} = a_i, \quad \forall i = 1,...,m && \text{(1.14b)}\\
& \sum_{i=1}^{m}\pi_{i,j} = b_j, \quad \forall j = 1,...,n && \text{(1.14c)}\\
& \pi_{i,j} \geq 0, \quad \forall i = 1,...,m, \quad j = 1,...,n && \text{(1.14d)}
\end{aligned}$$

In the context of numerical approximation to the OT problem, entropic regularization offers computational advantages, especially for large datasets. Specifically, entropy-regularization of the OT problem transforms the problem such that its solution is of a form easily written using matrix notation. This matrix, indeed the optimal transport plan for a prescribed regularization value, may be obtained using the Sinkhorn-Knopp algorithm (Sinkhorn 1964; Sinkhorn and Knopp 1967; Sinkhorn 1967) through a simple iterative procedure. It is worth noting here that the entropy term added to the objective function is scaled by a regularization coefficient (γ); for large values of this coefficient, the Sinkhorn OT plan is an increasingly diffused coupling, while for smaller values tending to zero, the Sinkhorn OT plan converges to that of the original OT problem (Cominetti and Martin 1994). Chapter 2 provides a discussion of these results prior to presenting algorithms that utilize the Sinkhorn OT problem to optimally reduce a large (support) set of data points to a smaller subset such that the information in these probability distributions is preserved to the greatest extent; in other words, the optimal subset obtained through these algorithms has the least Sinkhorn OT distance amongst all possible subsets.

1.2.2.3.2 Optimal transport between Gaussian mixtures

The entropy-regularized optimal transport (OT) problem presented in the previous section offers one method of obtaining OT plans, and consequently the OT distances, for large datasets. There have many other advances in the field of OT theory that have provided alternative methods for the same; one such method is the optimal transport between Gaussian mixtures presented by Chen *et al.* (2018). The detailed mathematical derivation for this OT variant is described in Chapters 3 and 4.

The optimal transport variant for transport between probability measures ζ and ω modeled as Gaussian mixtures $\mathbb{G}_\zeta$ and $\mathbb{G}_\omega$, respectively, builds upon the discrete Kantorovich OT framework, with conceptual modifications made to the measures' supports. Specifically, in this variant, the discrete supports available are modeled as finite-component Gaussian Mixture Models (GMMs). GMMs have been well-explored in literature for a number of applications (Roweis and Ghahramani 1999; Kostantinos 2000; McLachlan *et al.* 2019). Modeling the supports of the discrete measures as GMMs expresses the measures as linear weighted combinations of Gaussian measures $\left(\mathbb{G}_\zeta = \sum_{l=1}^{L_\zeta} w_l^\zeta \nu_l^\zeta, \mathbb{G}_\omega = \sum_{l'=1}^{L_\omega} w_{l'}^\omega \nu_{l'}^\omega\right)$. As a result, the OT problem between ζ and ω may be approximated by the OT problem between $\mathbb{G}_\zeta$ and $\mathbb{G}_\omega$. In the $\mathrm{OT}(\mathbb{G}_\zeta, \mathbb{G}_\omega)$ problem, the Gaussian components $(\nu_l^{L_\zeta}, \nu_{l'}^{L_\omega})$ are analogous to the Dirac point masses $(\delta_{x_i}, \delta_{y_j})$ in the original OT problem (Model 1.11), while the weighting proportions $(w_l^\zeta, w_{l'}^\omega)$ are akin to the weights (a_i, b_j), respectively. Therefore, the cost of transport between the measures is defined on the Gaussian point masses $(\nu_l^{L_\zeta}, \nu_{l'}^{L_\omega})$ and the probability conservation constraints are imposed on the weighting proportions $(w_l^\zeta, w_{l'}^\omega)$; in their work, Chen *et al.* (2018) defined this cost as the squared 2-Wasserstein distance. This cost is calculated using the closed-form expression for the W_2 distance between the Gaussian measure point masses developed by Takatsu (2011). The optimal transport problem between the Gaussian mixtures $\mathbb{G}_\zeta$ and $\mathbb{G}_\omega$ may be given as,

$$\begin{aligned}
z := \min_{\pi \in \Pi} \quad & \sum_{l=1}^{L_\zeta} \sum_{l'=1}^{L_\omega} c_{l,l'} \pi_{l,l'} && (1.15a)\\
\text{s.t.} \quad & \sum_{l'=1}^{L_\omega} \pi_{l,l'} = w_l^\zeta, \quad \forall l = 1, ..., L_\zeta && (1.15b)\\
& \sum_{l=1}^{L_\zeta} \pi_{l,l'} = w_{l'}^\omega, \quad \forall l' = 1, ..., L_\omega && (1.15c)\\
& \pi_{l,l'} \geq 0, \quad l = 1, ..., L_\zeta, \quad l' = 1, ..., L_\omega && (1.15d)
\end{aligned}$$

The optimal transport distance between GMMs from this model is the square root of the optimal objective value (z); this distance $(\sqrt{z})$ is proven to be a lower bound on $W_2(\zeta, \omega)$.

1.3 Thesis outline and structure

Having presented an overview of the background and theory pertinent to the problem statements tackled in this thesis, this section outlines the work flow and structure.

Chapter 2 addresses the problem of optimal scenario reduction and scenario tree generation using the entropy-regularized optimal transport variant. This work leverages the iterative numerical Sinkhorn algorithm for OT, and proposes algorithms for the reduction of large dimensional supersets of scenarios, to optimal smaller subsets with minimal loss in information for stochastic programming. This work provides a study of the proposed method's computational performance as well as solution quality, and highlights the efficacy of the algorithms proposed on case studies.

Chapter 3 leverages optimal transport between Gaussian mixture models (GMMs) to propose a novel ambiguity set construction method for distributionally robust optimization (DRO). This work is designed to reduce conservatism of the DRO solution by retaining the multimodal characteristics of the available data on process uncertainty, and further incorporating the same through first- and second-order moments into the GMM-OT-based ambiguity set construction step. Chapter 4 builds upon the DRO formulation proposed in Chapter 3 for distributionally robust chance-constrained programming (DRCCP). The studies performed show that the overconservativeness resulting from a distributionally robust treatment of the problem may be reduced by incorporating multimodal characteristics into the metric-based ambiguity set construction step, and demonstrate the application of the proposed formulations on case studies from financial and chemical engineering fields.

Chapter 5 pivots to the application of optimal transport theory to process monitoring. This work showcases the performance of the OT distance as a metric for change point detection in multivariate processes, and further gives a moving window approach to online process monitoring for fault detection. This work shows that the optimal transport distance provides good fault detection performance, and demonstrates its applicability on benchmark studies. Chapter 6 leverages the formulation proposed in Chapter 3 to assess the worst-case performance of a fault detection system under distributional ambiguity of multimodal processes (modeled as GMMs).

Chapter 2

Scenario Reduction and Scenario Tree Generation for Stochastic Programming using Entropy-Regularized Optimal Transport

Abstract: Scenario-based stochastic programming is a widely used method for optimization under uncertainty. The solution quality of this approach is dependent on the approximation of the underlying uncertainty distribution. Therefore, the optimal generation of scenarios (or scenario trees) is a pertinent research objective in stochastic programming. In this work, we approach the scenario reduction and scenario tree generation problem through the perspective of optimal transport, specifically entropy-regularized optimal transport. The availability of an iterative procedure to compute the optimal entropy-regularized transport plan between support sets, using the Sinkhorn-Knopp algorithm in lieu of conventional linear programming-based optimal transport, is found to decrease solution time appreciably, with a decrease in memory burden as well. We present algorithms for optimal scenario reduction and multistage scenario tree generation, and illustrate their use through two case studies. We show that the proposed approach generates high-quality scenarios whose use in stochastic programming offers solutions with good accuracy.

2.1 Introduction

Industrial processes generally operate under a varying range of uncertainties, from demand and supply, to resource availability and scheduling. Optimal decision-making, therefore, must take into account the underlying uncertainty in order to obtain practically feasible decisions. In this regard, it is imperative to assess the amount of distributional information about the uncertain parameters, if any, available for incorporation into optimization.

Two popular optimization approaches under uncertainty are robust optimization and stochastic programming. Robust optimization uses the support information about the underlying uncertainty in the optimization problem, and solves the problem for the worst case realization of uncertainty, thus offering a risk-averse solution. While this approach is often used in applications concerning process safety and design, robust optimization also leads to markedly higher costs. In contrast, stochastic programming, proposed by Dantzig (1955), comprises methods in which the probability distribution of the uncertain parameter is assumed to be known, or reasonably estimated from available process history. In order to solve the stochastic programming problem numerically, a deterministic equivalent of the problem may be obtained through scenario-based approximation method. In this method, the optimization problem is solved for a finite set of scenarios generated to represent the uncertainty in discrete realizations with corresponding probabilities of occurrence. Generally, a large number of scenarios would capture the true uncertainty distribution better than a smaller number. However, it is found that as the number of scenarios increases, the problem complexity, as well as solution time, increases greatly. Therefore, there is a need to explore methods of optimal scenario reduction.

The concept of scenario reduction was pioneered by Dupačová *et al.* (2003), who proposed the use of a natural probability metric as an approximation metric in order to obtain the closest subset from a larger superset of scenarios. The authors proposed the use of Fortet-Mourier metrics for the reduction of electrical load scenario trees used for power management under uncertainty. The stability of multistage stochastic programs was analyzed by Heitsch *et al.* (2006) who posited that the reduction of multistage scenario trees should be based on L_r distances as well as filtration distances. Subsequently, Heitsch and Römisch (2009a) derived algorithms for the reduction of multistage scenario trees using this idea. Xu *et al.* (2012) developed algorithms using K-means clustering and LP moment-matching methods to approximate multistage scenario trees from a large scenario fan description of the uncertain parameter. In this method, the authors generated a multistage scenario tree from a scenario fan composed of a large number of profiles evolving in time, generated as a result of random processes. At each stage, the superset of points is reduced to a smaller subset comprising cluster centers chosen by the K-means clustering technique. Chen and Yan (2018) designed a scenario tree reduction algorithm through clustering tree nodes based on a new distance function to measure the difference between two scenario trees.

Li and Floudas (2014) treated the scenario reduction problem as a mixed integer linear programming (MILP) problem. The authors designed an MILP-based method for scenario reduction using Kantorovich Distance. In this method, the binary variables denote whether the scenario is retained or removed to meet the new reduced scenario set size. The constraints in the MILP model pertain to the conservation of probability mass from the superset to the reduced subset. This is a single-step approach in the sense that this optimization problem is solved once, and the optimal reduced scenario subset is obtained. In contrast, Li and Li (2016) proposed a computationally efficient iterative algorithm for scenario reduction that circumvents the limitations posed by the aforementioned MILP formulation through the use of transportation distance. In a recent work by Zhou *et al.* (2019), the authors introduce an improved method using K-means with a typicality degree approach for scenario reduction.

In addition to transportation metric-based techniques, scenario reduction has also been accomplished for decision-making under uncertainty using other methods. Karuppiah *et al.* (2010) proposed a heuristic scenario selection strategy that the overall probability of occurrence of a particular realization of any uncertain parameter in the final set of scenarios should be equal to the probability of the uncertain parameter taking on that particular value. Meira *et al.* (2016) performed scenario reduction using representative models in oil fields. Arpón *et al.* (2018) have chosen the Conditional-Value-at-Risk (cVaR) measure as the objective function in order to accomplish the reduction. In the work of Hu and Li (2019), they balanced the objective of scenario reduction by maximizing the likeness between the original scenario superset and the reduced subset, while simultaneously minimizing the correlation loss before and after the reduction process. Silvente *et al.* (2019) proposed a scenario tree reduction method using sensitivity analysis for nonlinear optimization models. In their work, the authors chose the sensitivity of scenarios as a measure of identifying which scenarios to retain in a larger superset. Scenario reduction has also been approached through the purview of machine learning. Li and Gao (2019) used deep learning to reduce scenario supersets. In their work, the authors transformed scenarios into a format similar to that of images, and fed these transformations to a deep convolutional neural network to obtain a similar image-like output of the reduced subset. Medina-Gonzalez *et al.* (2020) proposed a scenario-reduction method that integrates data mining, graph theory and community detection concepts to represent the uncertain information as a network and identify the most efficient communities/clusters. Bounitsis *et al.* (2022) proposed a data-driven MILP model for the distribution matching problem behind scenario reduction. They have shown improved efficiency through integration of copula-based simulation and clustering method.

In this work, we aim to improve the LP-based scenario reduction strategy, proposed by Li and Li (2016) by replacing the inner linear optimal transport problem with entropy-regularized optimal transport solved using the Sinkhorn-Knopp algorithm (Section 2.3). We present the algorithm for optimal scenario reduction using Sinkhorn distance (Section 2.4), and extend its application to the optimal generation of multistage scenario trees for stochastic programming (Section 2.5). Furthermore, we use the process distance metric (Pflug and Pichler 2012) to evaluate the approximation quality between the generated scenario tree, and the original scenario fan. Finally, we illustrate the use of these algorithms on two case studies (Section 2.6).

2.2 Optimal transport distance between distributions

The extent of similarity between two probability distributions is quantified in literature through a number of measures, such as Kullback-Leibler Divergence (KLD), Hellinger Distance, power distance, and correlation similarities. However, there are certain instances where measures like KLD cannot be used, as is the case when the level of similarity between two distributions defined on different probability spaces is to be quantified.

When distributions defined on different support sets are to be compared, the problem can be viewed through the lens of optimal transport. Optimal transport was first studied by Monge (1781) in the context of transportation of mined soil from the quarry to various construction sites, famously titled the “earth-mover’s problem” in literature. The optimal transport problem aims to find the most efficient manner in which to transfer probability mass from the elements of one distribution to another, while minimizing a chosen cost function. It follows intuitively that if the two distributions in question are vastly dissimilar, the task of moving the probability masses from one distribution to the other requires more effort, which is quantified by the optimal transport distance.

In the pioneering work of Monge (1781), a continuous transport map $f(x)$ was estimated between two probability distributions $\mathbb{P}(x)$ and $\mathbb{Q}(y)$, with support sets X and Y, respectively, so as to minimize the expected cost of transportation $c(x, f(x))$, as given by Equation 2.1,

$$M(\mathbb{P}, \mathbb{Q}) = \inf_{f \in MP} \int_X c(x, f(x)) d\mathbb{P}(x) \tag{2.1}$$

MP denotes the set of all mappings that preserve the transfer of probability from $\mathbb{P}(x)$ to $\mathbb{Q}(y)$ for any Borel subset A of Y (Equation 2.2).

$$MP = \left\{ f : X \to Y \middle| \int_{f^{-1}(A)} d\mathbb{P}(x) = \int_A d\mathbb{Q}(y) \right\} \tag{2.2}$$

It must be noted that the Monge formulation of the optimal transport problem does not allow for probability masses to be split during transfer.

2.2.1 Kantorovich Distance and Wasserstein Distance

Kantorovich (1942) approached the optimal transport problem through a relaxed formulation, allowing for the splitting for probability masses; this formulation uses the concept of a transport plan which is given by (Equation 2.3),

$$K(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{X \times Y} c(x, y) d\pi(x, y) \tag{2.3}$$

Here, $\Pi(\mathbb{P}, \mathbb{Q})$ is the set of all joint distributions whose marginal distributions are $\mathbb{P}(x)$ and $\mathbb{Q}(y)$.

The optimal transport problem can be formulated as a linear programming (LP) problem for transport between discrete distributions. Consider a simplified version of the same (Model 2.4),

$$\begin{aligned} \text{KD} = \min_{\pi_{i,j}} \quad & \sum_{i,j} c_{i,j}\pi_{i,j} && \text{(2.4a)} \\ \text{s.t.} \quad & \sum_{j} \pi_{i,j} = a_i \quad \forall i && \text{(2.4b)} \\ & \sum_{i} \pi_{i,j} = b_j \quad \forall j && \text{(2.4c)} \\ & \pi_{i,j} \geq 0 \qquad \forall i,j && \text{(2.4d)} \end{aligned}$$

In Model 2.4, the variables $\pi_{i,j}$ are elements of the optimal transport plan, and represent the amount of probability mass that is transferred from one distribution's elements ($x_i \sim \mathbb{P}(x)$) to those of another distribution's ($y_j \sim \mathbb{P}(y)$). The cost of moving probability mass between elements x_i and y_j is computed as $c_{i,j}$, which is often computed using the norm $c(x,y) = \|x-y\|$. For example, 1-norm and 2-norm based distance for two $n-$dimensional points x and y is given as: $\|x-y\|_1 = \sum_{k=1}^{n} |x_k - y_k|$, and $\|x-y\|_2 = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$.

The objective function in Equation 2.4a represents the total cost in moving probability masses from $\mathbb{P}(x)$ to $\mathbb{Q}(y)$ to be minimized for optimal transport. The constraints in Model 2.4 represent the conservation of probability mass with respect to $\mathbb{P}(x)$ and $\mathbb{Q}P(y)$, respectively, subject to non-negativity constraints on $\pi_{i,j}$. The vectors a and b in Equations 2.4b and 2.4c contain the probability masses of the elements $x_i \sim \mathbb{P}(x)$ and $y_j \sim \mathbb{Q}(y)$, respectively. It must be noted that the transport plan obtained through the Kantorovich formulation of the optimal transport problem is not a one-to-one mapping, ergo allowing for the probability mass of one discrete element in the source distribution to be split across that of multiple elements in the destination.

The above optimal transport problem formulation is a special case of the p-Wasserstein distance ($p \geq 1$), which is the optimal objective of the following optimal transport problem (Equation 2.5),

$$W_p(\mathbb{P},\mathbb{Q}) = \left(\inf_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \int_{X \times Y} c(x,y)^p d\pi(x,y) \right)^{\frac{1}{p}} \tag{2.5}$$

For $p = 1$, the 1-Wasserstein distance is termed as the Kantorovich Distance (KD), which is the optimal objective of the optimal transport problem in Model 2.4.

2.3 Entropic regularization and Sinkhorn distance

The optimal transport plan $\pi_{i,j}$ in Model (2.4) is an $M \times N$-dimensional decision variable, where M and N are the sizes of the support sets X and Y. For large dimensional optimal transport problems, the formulation in Model 2.4 encounters a memory bottleneck for computation of the optimal transport plan . To mitigate this issue, Lellmann *et al.* (2014) used the concept of Kantorovich-Rubenstein duality in which the dual formulation of the optimal transport problem is solved. Another method to tackle the memory burden in conventional optimal transport makes use of entropy regularization of the problem(Cuturi 2013) that be solved using an iterative approach, rather than a linear programming-based formulation.

A matrix with lower entropy has most of its non-zero values concentrated in fewer points, while a matrix with higher entropy is smoother and has a more uniform distribution of its non-zero values across its elements. Therefore, entropy regularization is carried out to make the coupling matrix smoother by introducing the regularization coefficient (γ) and the entropy of the transport plan ($H = -\sum_{i,j} \pi_{i,j}[\log(\pi_{i,j}) - 1]$) to the objective function(Peyré, Cuturi, *et al.* 2019). Notice that the objective is of minimization type and we use negative entropy such that the entropy can be maximized to achieve smooth transport plan. The entropy-regularized optimal mass transport problem is given as Model 2.6,

$$\text{SD} = \min_{\pi_{i,j}} \quad \sum_{i,j} c_{i,j}\pi_{i,j} + \gamma\pi_{i,j}[\log(\pi_{i,j}) - 1] \tag{2.6a}$$

$$\text{s.t.} \quad \sum_{j} \pi_{i,j} = a_i \qquad \forall i \tag{2.6b}$$

$$\sum_{i} \pi_{i,j} = b_j \qquad \forall j \tag{2.6c}$$

$$\pi_{i,j} \geq 0 \qquad \forall i, j \tag{2.6d}$$

Note that the Sinkhorn distance is not a metric since it is biased: $SD(\mathbb{P}, \mathbb{P}) \neq 0$.

The smoothness of the coupling matrix increases with the value of γ. Therefore, lower the value of γ, closer the solution to that of the original optimal transport problem(Peyré, Cuturi, *et al.* 2019). The entropy-regularized transport problem can be solved using the Sinkhorn-Knopp algorithm (Peyré, Cuturi, *et al.* 2019; Sinkhorn 1964).

The Lagrangian for Model (2.6) is given as (Equation 2.7),

$$L = \left\{ \sum_{i,j} c_{i,j}\pi_{i,j} + \gamma \sum_{i,j} \pi_{i,j}[\log(\pi_{i,j}) - 1] \right\} + \boldsymbol{\alpha}^{\intercal}[\mathbf{\Pi}(\mathbf{1}_{N\times 1}) - a] + \boldsymbol{\beta}^{\intercal}[\mathbf{\Pi}^{\intercal}(\mathbf{1}_{M\times 1}) - b] \tag{2.7}$$

where $\boldsymbol{\alpha}_{M\times 1}$ and $\boldsymbol{\beta}_{N\times 1}$ are Lagrange multiplier vectors corresponding to constraints 2.6b and 2.6c, respectively, and $\mathbf{\Pi}$ is the matrix $[\pi_{i,j}]$. Based on stationary conditions, differentiating the Lagrangian with respect to the transport plan, we have (Equation 2.8),

$$\frac{\partial L}{\partial \pi_{i,j}} = c_{i,j} + \gamma \log(\pi_{i,j}) + \alpha_i + \beta_j = 0 \tag{2.8}$$

Solving for $\pi_{i,j}$, the expression for the optimal transport plan is obtained as (Equation 2.9),

$$\underbrace{\pi_{i,j}}_{\mathbf{\Pi}} = \underbrace{\exp\left(-\frac{\alpha_i}{\gamma}\right)}_{\boldsymbol{U}} \underbrace{\exp\left(-\frac{c_{i,j}}{\gamma}\right)}_{\boldsymbol{M}} \underbrace{\exp\left(-\frac{\beta_j}{\gamma}\right)}_{\boldsymbol{V}} \tag{2.9}$$

When the entries of the optimal transport plan are viewed in matrix form, Equation 2.9 shows that the optimal transport plan $\mathbf{\Pi}$ is obtained from a positive kernel cost matrix $\boldsymbol{M}$ scaled by diagonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$. For a matrix $\boldsymbol{M}$ with positive entries, there exist two diagonal matrices, $\boldsymbol{U}$ and $\boldsymbol{V}$, with positive diagonal entries such that the matrix product $\boldsymbol{UMV}$ has the i^{th} row sum a_i and j^{th} column sum b_j - this is termed as the matrix scaling problem (Nemirovski and Rothblum 1999). The corresponding iterative algorithm for computing these vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ (constructed using diagonal entries of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively) is given as (Equation 2.10),

$$\boldsymbol{u}^{(g+1)} = \frac{\boldsymbol{a}}{\boldsymbol{M}\boldsymbol{v}^{(g)}} \qquad \boldsymbol{v}^{(g+1)} = \frac{\boldsymbol{b}}{\boldsymbol{M}^{\intercal}\boldsymbol{u}^{(g+1)}} \tag{2.10}$$

Here, g represents the iterations in the calculation of $\boldsymbol{u}$ and $\boldsymbol{v}$. The proof of convergence of this iterative procedure (Deming and Stephan 1940; Bacharach 1965) was worked upon by (Sinkhorn 1964). The errors associated with the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are given by (Equation 2.11),

$$\epsilon_a = \frac{||\boldsymbol{u}\boldsymbol{M}\boldsymbol{v} - \boldsymbol{a}||}{||\boldsymbol{a}||} \qquad \epsilon_b = \frac{||\boldsymbol{v}\boldsymbol{M}^{\intercal}\boldsymbol{u} - \boldsymbol{b}||}{||\boldsymbol{b}||} \tag{2.11}$$

The vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are iteratively computed until both error terms ϵ_a and ϵ_b are below the specified error threshold, $\bar{\epsilon}$. The entries of the optimal transport plan $\pi_{i,j} \in \Pi$ are calculated using Equation 2.9, where the final $\boldsymbol{u}$ and $\boldsymbol{v}$ vectors comprise the principal diagonal elements of matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Finally, the Sinkhorn Distance (SD) is computed as (Equation 2.12),

$$\text{SD} = \sum_{i,j} c_{i,j}\pi_{i,j} + \gamma \sum_{i,j} \pi_{i,j}[\log(\pi_{i,j}) - 1] \tag{2.12}$$

A step-wise procedure for the calculation of Sinkhorn distance between two sets I and S is outlined in Algorithm 1.

Algorithm 1: Sinkhorn Algorithm for Optimal Transport with Entropic Regularization

```
Input: Set $I = \{x_i\}$, with probability $a_i$ for element $x_i \in X, x_i \sim \mathbb{P}(x)$
       Set $S = \{y_j\}$, with probability $b_j$ for element $y_j \in Y, y_j \sim \mathbb{Q}(y)$
       Error threshold $\bar{\epsilon}$
Procedure:
Step 1: Evaluate cost matrix entries $c_{i,j} := \|x_i - y_j\|_2$
Step 2: Evaluate kernel function entries $m_{i,j} := \exp(-\frac{c_{i,j}}{\gamma})$
Step 3: Initialize iteration: $g \leftarrow 1$
        Initialize vector $\boldsymbol{v}^{(g)}$ (value 1 for all entries)
while $\epsilon_a, \epsilon_b \geq \bar{\epsilon}$ do
    Evaluate vectors $\boldsymbol{u}^{(g+1)} = \frac{\boldsymbol{a}}{\boldsymbol{M}\boldsymbol{v}^{(g)}}$, $\boldsymbol{v}^{(g+1)} = \frac{\boldsymbol{b}}{\boldsymbol{M}^\top \boldsymbol{u}^{(g+1)}}$
    Evaluate errors $\epsilon_a = \frac{\|\boldsymbol{u}^{(g+1)}\boldsymbol{M}\boldsymbol{v}^{(g+1)} - \boldsymbol{a}\|}{\|\boldsymbol{a}\|}$, $\epsilon_b = \frac{\|\boldsymbol{v}^{(g+1)}\boldsymbol{M}^\top \boldsymbol{u}^{(g+1)} - \boldsymbol{b}\|}{\|\boldsymbol{b}\|}$
    $g \leftarrow g + 1$
end
Step 4: Calculate optimal transport plan $\mathbf{\Pi} = [\pi_{i,j}] = \boldsymbol{UMV}$
Step 5: Calculate Sinkhorn distance
            $\mathrm{SD} = \sum_{i,j} c_{i,j}\pi_{i,j} + \gamma \sum_{i,j} \pi_{i,j}[\log(\pi_{i,j}) - 1]$
Output: Sinkhorn distance SD
        Optimal transport plan $\pi_{i,j}$
```

Next, we illustrate an instance of entropy-regularized (SD-based) optimal transport. Here, the optimal transport problem entails the movement of probability point masses of one distribution, represented by the blue dots in the outer annulus, to those of another distribution, represented by the red dots in the inner spiral, in Figure 2.1, using the Sinkhorn distance iterative calculations detailed in Algorithm 1.

Figure 2.1 illustrates the effect of the regularization coefficient γ on the smoothness of the optimal transport plan. The solid lines represent maximum probability transfers, while the dotted lines represent probability mass transfer greater than 5%. Here, the optimal transport problem was solved for four values of $\gamma = 10, 1, 0.1$ and 0.01. It is observed that as the value of γ decreases, the smoothness of the optimal transport plan reported decreases as well. This is illustrated by the gradual disappearance of the dotted lines as γ increases, thereby showing that the optimal transport plan at lower values of γ mirrors the plan obtained from conventional optimal transport

(γ = 0), and contains a number of zero values.

Choice of regularization coefficient (γ): In the entropy-regularized optimal transport problem, γ is introduced to smoothen the optimal transport plan such that the number of non-zero elements reduces with an increase in γ. In practice, γ is chosen to be sufficiently small so as to best approximate the conventional optimal transport plan, while large enough to avoid numerical issues in the computation of the exponential terms involved in the Sinkhorn algorithm (Algorithm 1).

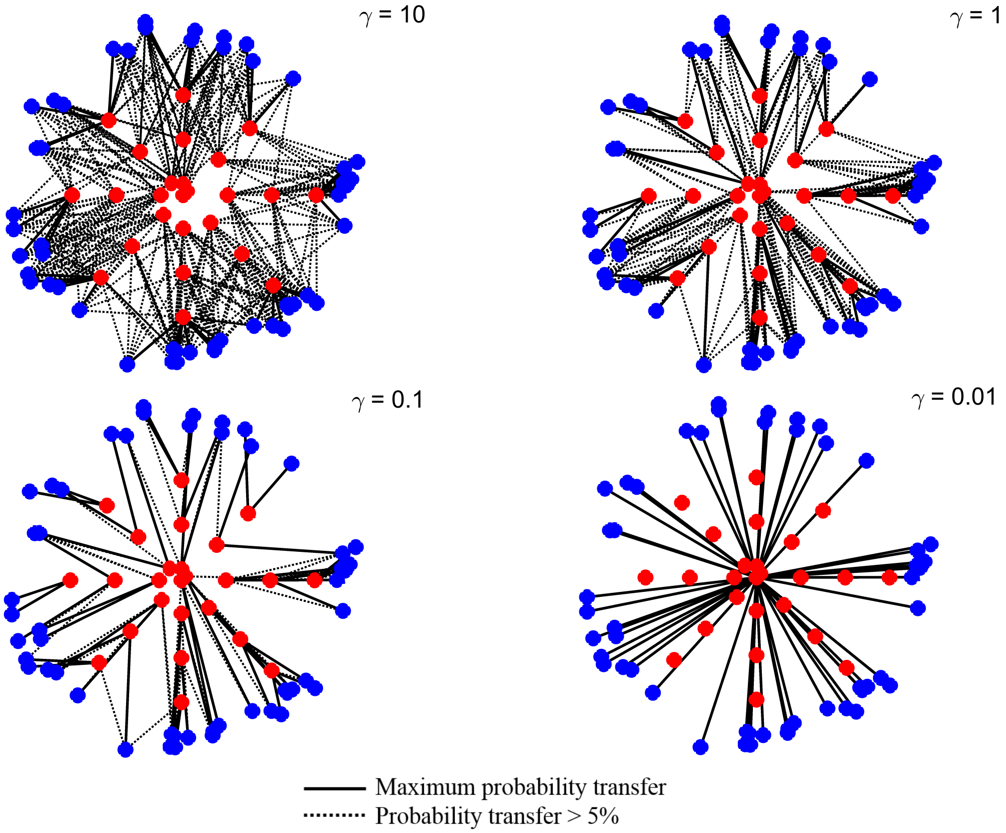


Figure 2.1: Effect of the regularization coefficient (γ) on smoothness of the transport plan

2.4 Scenario reduction using Sinkhorn distance

In this section, we propose an algorithm for optimal scenario reduction using Sinkhorn distance for a large superset of points, denoted by I, to obtain a reduced subset S using an iterative method where the extent of similarity between I and S is computed using the Sinkhorn distance. Section 2.4.1 presents the problem through a mixed integer nonlinear optimization (MINLP) formulation. Section 2.4.2 presents an iterative solution algorithm for this problem using the Sinkhorn distance calculation detailed in Algorithm 1, followed by a numerical illustration of the algorithm.

2.4.1 MINLP model

The mixed integer linear programming (MILP) model presented in Li and Floudas (2014) may be used to reduce a superset of scenarios to an optimal reduced subset. This optimization model minimizes the Kantorovich Distance (KD) between the superset and subset. In a similar manner, a mixed integer nonlinear programming (MINLP) formulation of the entropy regularized optimal transport problem may be developed, as shown in Model 2.13.

$$
\begin{aligned}
\min_{\pi_{i,i'},b_{i'},y_i} \quad & \sum_{i\in I}\sum_{i'\in I}\pi_{i,i'}c_{i,i'} + \sum_{i\in I}\sum_{i'\in I}\gamma\pi_{i,i'}[\log(\pi_{i,i'}) - 1] && && (2.13a)\\
\text{s.t.} \quad & \sum_{i'}\pi_{i,i'} = a_i && \forall i \in I && (2.13b)\\
& \sum_{i\in I}\pi_{i,i'} = b_{i'} && \forall i' \in I && (2.13c)\\
& \sum_{i'\in I} b_{i'} = 1 && && (2.13d)\\
& \sum_{i\in I} y_i = r && && (2.13e)\\
& \epsilon y_i \le b_i \le y_i && \forall i \in I && (2.13f)\\
& \pi_{i,i'} \ge 0 && \forall i, i' \in I && (2.13g)\\
& b_{i'} \ge 0 && \forall i' \in I && (2.13h)\\
& y_i \in \{0,1\} && \forall i \in I && (2.13i)
\end{aligned}
$$

In this model, r is the desired number of scenario is the reduced set, the binary variable y_i denotes the status of retention of the scenario $i \in I$ in the optimal subset. The variable b_i denotes the amount of probability mass assigned to the retained scenario i. Finally, $\pi_{i,i'}$ denotes the amount of probability mass transported between the elements $i, i' \in I$. The parameter ϵ is taken to be a small value of the order of magnitude 10^{-9}, in order to define the relationship between the binary status variable y_i and assigned probability mass b_i.

In addition to utilizing a nonlinear objective function to be minimized, Model 2.13 contains binary variables, whose number increases as the size of the superset to be reduced increases, thus escalating the problem complexity and computational time involved. This issue of rapidly surging problem complexity, with the number of integer variables, presents a clear need for an iterative solution of a simplified optimization problem. It is also noted that an MILP formulation for scenario reduction based on KD can be easily obtained after removing the second term in the objective function.

2.4.2 Solution algorithm

The Sinkhorn distance calculation presented in Algorithm 1 details the procedure for transporting probability masses of the elements of one distribution (with finite support set I) to the other distribution (with support set S). Therefore, SD may be viewed as a quantitative measure of dissimilarity between the two distributions: a smaller value of SD is indicative of more similar probability distributions. This interpretation of the optimal transport problem can be extended to finding the optimal S^* such that $\text{SD}(I, S)$ is as small as possible. To this end, we propose the following iterative procedure for optimal scenario reduction in Algorithm 2.

The scenario superset I is fed as input to this algorithm, and it is to be reduced to an optimal subset S^* containing r scenarios. At every iteration (d), the Sinkhorn

Algorithm 2: Optimal scenario reduction using Sinkhorn distance

```
Input: Superset $I$, Subset size $r$, Error threshold $\beta_{\mathrm{SD}}$
Procedure: Initialize iteration ($d \leftarrow 1$)
while $RelativeError \geq \beta_{SD}$ do
    Step 1: Define subset ($S$)
    if $d = 1$ then
        $S^{(d)}$ is set as $r$ randomly chosen points in $I$
    else
        $S^{(d)} := S^{(d-1)}$
    end
    Step 2: Calculate Sinkhorn distance ($\mathrm{SD}^{(d)}$) and optimal transport plan
      ($\Pi^{(d)}$) using Algorithm 1
    Step 3: Form clusters $C_1, C_2, ..., C_r$ around points in $S$, using maximum
      probability transfer links in $\Pi^{(d)}$
    Step 4: Calculate relative error
    if $d = 1$ then
        Proceed to step 5
    else
        RelativeError = $\frac{\mathrm{SD}^{(d-1)} - \mathrm{SD}^{(d)}}{\mathrm{SD}^{(d-1)}}$
    end
    Step 5: Update cluster centers $S^{(d)}$ as points with minimum in-cluster
      transportation distance

    $d \leftarrow d + 1$
end
Step 6: Obtain final reduced subset $S^* := S^{(d)}$, and corresponding clusters
  $C_1, C_2, ...C_r$
Step 7: Compute probabilities of points in optimal reduced subset $S^*$ as

    $$p_r = \frac{\text{Number of points in cluster } C_r}{\text{Number of points in superset } I}$$

Output: Final reduced subset $S^*$ containing $r$ points, the corresponding
  probability and clusters $C_1, C_2, ..., C_r$
```

distance between I and $S^{(d)}$ is computed using Algorithm 1. Using the maximum probability transfer links from the optimal transport plan $\Pi^{(d)}$, clusters are formed around the points retained in $S^{(d)}$. The relative error between the Sinkhorn distance values at successive iterations is computed; this value is compared with the specified threshold value (β_{SD}) as the termination criterion. If the computed relative error is above the threshold, the cluster centers are updated to give minimum in-cluster transportation distance, and this updated subset is used in the following iteration. The optimal subset S^* is obtained from the final iteration of the algorithm. In this work, we assume that all points in the superset have an equal probability of occurrence. The probabilities of the optimally retained scenarios in S^* are computed using the cardinality of the optimal clusters. The optimal subset S^* with the corresponding probability vector, and the final clusters $C_1, C_2, ..., C_r$, are reported as outputs. Note that each cluster contains the values as well as the indices of the points with respect to the superset I.

Choosing number of scenarios (r) in the optimally reduced subset

In the scenario reduction problem addressed in this work, the number of scenarios (r) in the optimally reduced subset is a fixed, user-defined parameter. Ideally, through the law of large numbers, a large r ensures that the optimal solution to the scenario-based problem tends to the true optimal solution. However, too large a value of r might make the problem computationally difficult to solve, and therefore, the user may compute a large enough value of r. Kleywegt *et al.* (2002) proved that a large enough value of r may be chosen as,

$$r \geq \frac{1}{\gamma(\delta, \epsilon)} \log\Big(\frac{|S \backslash S^{\epsilon}|}{\alpha}\Big) \tag{2.14}$$

Here, $1 - \alpha$ is the user-defined confidence level for an $\epsilon-$optimal solution where $\epsilon \geq 0, 0 \leq \delta \leq \epsilon$, and $\gamma(\delta, \epsilon) \coloneqq \ \min_{x \in S \backslash S^{\epsilon}} I_x(-\delta)$, $I(.)$ being the exponential rate function. In the event that the above expression is difficult to compute, Kleywegt *et al.* (2002)

have proposed an algorithm that allows for a dynamic adjustment of r through an iterative process.

Illustrating example

In this section, we present a numerical illustrative example of the proposed optimal scenario reduction algorithm using Sinkhorn distance (Algorithm 2) presented in Section 2.4.2. Here, a superset I comprising 20 points (blue) is to be reduced to a subset containing 3 points. Initially, a random subset S containing 3 points (red) is chosen (Figure 2.2a). The Sinkhorn distance between I and S is calculated, using Algorithm 1, as 1.3379, and links (black) are established between points in the superset and subset, representing the clusters formed using the optimal transport plan. The subset S is updated (S^{new}) as the points in each cluster that give the minimum in-cluster transportation distance. The SD between I and S^{new} is calculated to be 0.9824 (Figure 2.2b). The relative error in Sinkhorn distance is found to be 26.57%. In this example, the chosen threshold (β_{SD}) for relative error is 5%, that is, the clusters are successively updated until the relative error in Sinkhorn distance is below 5%. In the next iteration, the SD between I and the updated S^{new} is 0.9184, and the new relative error in SD is 6.52%. In the following iteration, it is observed that the updated subset does not change from that of the previous iteration; here, the SD remains 0.9184 and the relative error in SD is, thus, 0%. Therefore, the subset from the final iteration is chosen as the optimal subset of scenarios (Figure 2.2c).

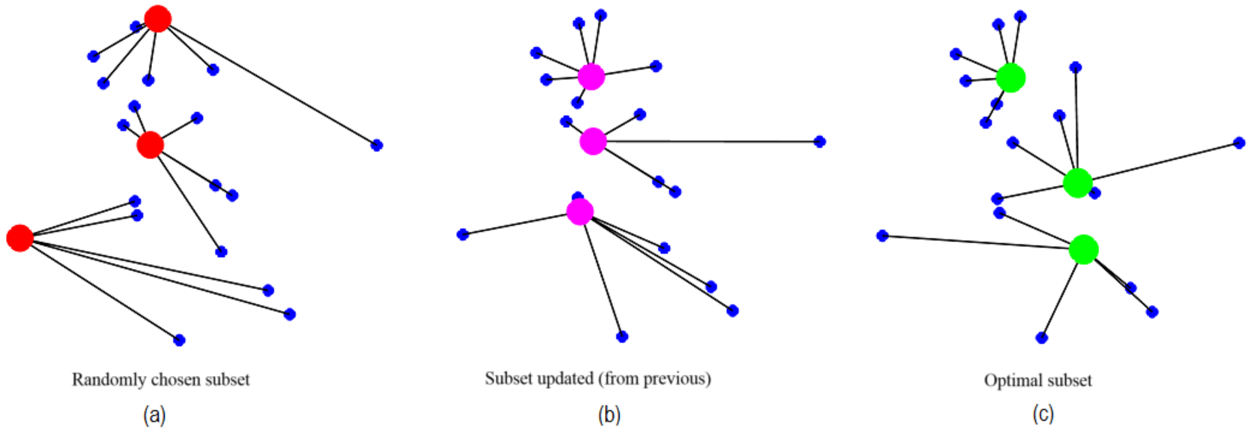


Figure 2.2: Illustration: optimal scenario reduction using Sinkhorn distance

2.4.3 Computational study

Figure 2.3 depicts the evolution in computational time for scenario reduction of increasingly large superset sizes, from 1000 to 20000 scenarios, to an optimal subset containing 50 scenarios. The figure compares the results from the proposed SD-based reduction in this work, to the LP-based scenario reduction proposed by Li and Li (2016). The LP-based reduction is found to be computationally slower than the proposed SD-based approach. Both scenario reduction schemes were tested on a desktop computer running on Intel(R) Core(TM) i5 CPU @3.2GHz, 4 cores with 8 GB of physical RAM.

Figure 2.4 compares the Sinkhorn distance between a superset of 100 normally distributed scenarios to a number of reduced size subsets obtained through the MINLP approach in Model 2.13, and the proposed iterative SD-based algorithm (Algorithm 2). Since the iterative approach utilizes a random starting subset to begin the scenario reduction process, 25 trials were conducted to obtain an average SD for this approach, as depicted by the boxplot. The numerical results for Figure 2.4 are presented in Table 2.1. The MINLP formulation runs to the default GAMS resource limit of 1000 seconds, while each trial of the proposed iterative SD-based algorithm

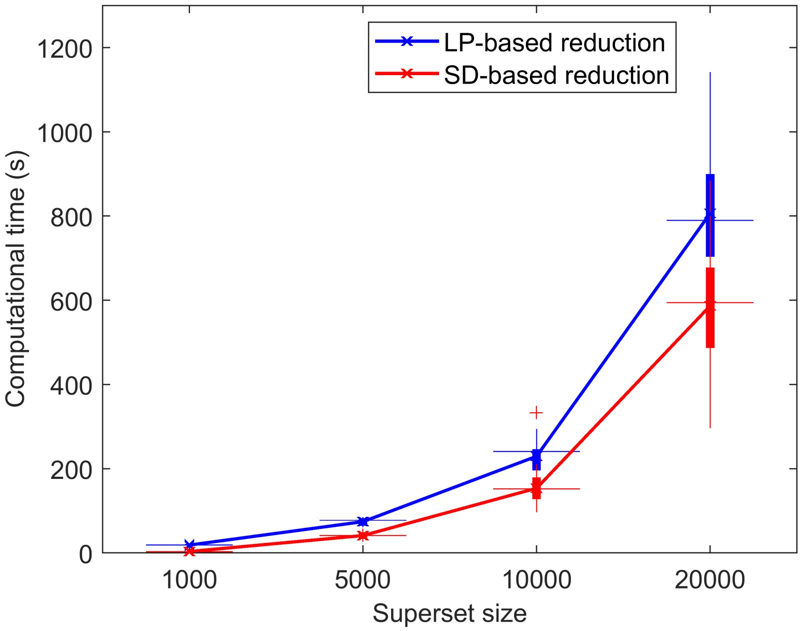


Figure 2.3: A comparison of computational times using the LP-based algorithm (Li and Li 2016), and the proposed SD-based approach

takes less than 10 seconds. It must further be noted that the MINLP formulation was run on GAMS using the BONMIN (Bonami and Lee 2011) solver, which reports the locally optimal solution.

Table 2.1: A comparison of evolution of Sinkhorn distance for the MINLP approach versus the proposed iterative SD-based algorithm for scenario reduction

Size of	Size of	SD	Mean SD
superset	reduced subset	(MINLP approach)	(proposed iterative algorithm)
100	20	0.2686	0.2580
	40	0.1061	0.1601
	60	0.0180	0.1178
	80	0	0.0636
	95	0	0.0091

Table 2.2 presents a summary of different scenario reduction approaches through

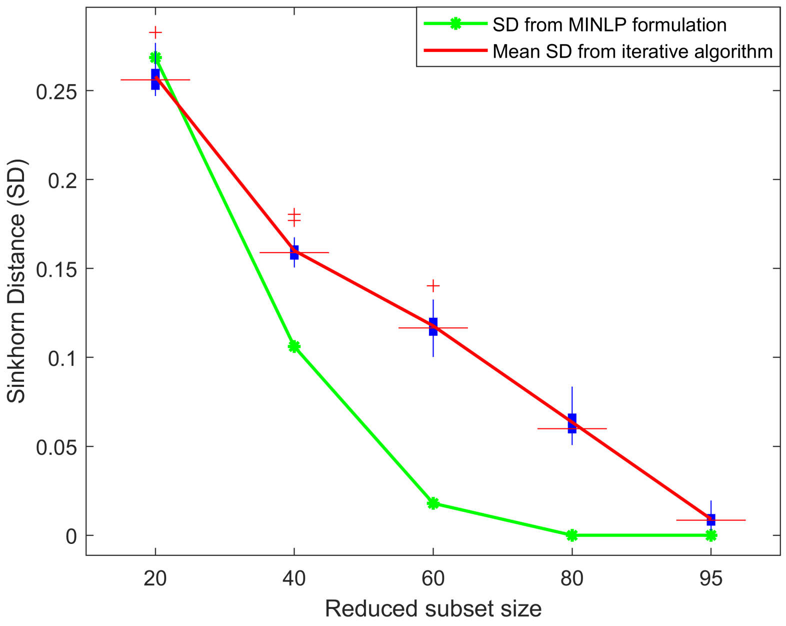


Figure 2.4: A comparison of Sinkhorn distance between a 100 scenario superset and reduced size subsets via the MINLP approach and the proposed SD-based algorithm for scenario reduction

the lens of optimal transport. The entropy-regularized formulation of the optimal transport problem enjoys a computational time advantage due to the availability of the Sinkhorn-Knopp iterative algorithm to obtain a smoother optimal transport plan, as compared to the conventional formulation, especially when the datasets involved are high-dimensional in nature. The exact scenario reduction problem utilizes binary variables that depict the status of retention or removal of scenarios from the superset; in either the conventional, or the entropy-regularized formulations, the optimization problem suffers from a slow solution process. To this end, an iterative LP-based algorithm was developed by Li and Li (2016). The iterative Sinkhorn distance-based approach to scenario reduction presented in this section is shown to perform faster than the iterative LP-based method (Li and Li 2016), as shown in Figure 2.3.

Table 2.2: A summary of scenario reduction approaches through optimal transport

	Kantorovich Distance	**Sinkhorn Distance**
	(KD)	**(SD)**
	Conventional optimal transport	*Entropy-regularized optimal transport*
Distance evaluation problem	LP problem [Model 2.4]	NLP problem with convex objective and linear constraints [Model 2.6] Efficiently solved using numerical iterations
Exact scenario reduction problem	MILP problem Li and Floudas (2014)	MINLP problem with convex objective and mixed integer linear constraints [Model 2.13]
Iterative scenario reduction algorithm	- Integer variables are removed - Each iteration solves an LP formulation of the conventional optimal transport Li and Li (2016)	- Integer variables are removed - Each iteration solves a numerical computation of the entropy-regularized optimal transport problem [Algorithm 2]

2.5 Multistage scenario tree generation algorithm

In this section, we extend the scenario reduction algorithm presented in Section 2.4.2 to the problem of optimal scenario tree generation from a scenario fan. A scenario fan comprises a number of profiles generated as a result of stochastic processes and the task is to obtain a multistage scenario tree to be used for stochastic programming. The algorithm presented in this section uses the procedure outlined in Algorithm 2 in a stage-wise manner to obtain the nodes of the scenario tree. Section 2.5.1 describes the proposed algorithm for optimal scenario tree generation, while Section 2.5.2 introduces the Process Distance (PD) measure that is used to evaluate the approximation quality of scenario trees generated using the proposed algorithm. A process distance-based solution quality analysis of the scenario trees obtained through the proposed algorithm is also presented.

2.5.1 Proposed workflow

The generation of a multistage scenario tree from a scenario fan is treated as a sequential extension of the scenario reduction problem, as described in Section 2.4.2.

Here, a scenario fan is considered to be a collection of profiles generated via stochastic processes over time, originating from a single root. An important convention to be clarified in this work is that of the "stage".

In this work, we refer to the root node of the scenario tree as "stage 0". The subsequent stages are numbered 1, 2, ..., and so on. With this notation, a two-stage stochastic programming problem is solved using a scenario tree with stage 0 and stage 1. Similarly, a n-stage stochastic programming problem is solved using a scenario tree with stages 0, 1, ..., $n-1$.

The proposed algorithm 3 for optimal scenario tree generation contains two main loops:

- Time stages ($t = 1, 2, ..., T$),
- Nodes over each stage ($w = 1, 2, ..., n_{t-1}$)

Let $B_t, t = 1, 2, ..., T$ describe the number of leaf nodes that originate out of nodes at stage t. The total number of leaf nodes n_t at each stage t is $n_t = n_{t-1}B_t$. Here, $n_0 = 1$ by default. At each stage t, an optimal set of nodes approximating the superset of nodes at the time step pertaining to that stage in the scenario fan is found. However, the number of such supersets to reduce at each stage depends on the number of nodes present in the previous stages. At each stage $t \geq 1$, there exist multiple supersets to be reduced, denoted by I_w^t, where $w = 1, 2, ..., n_{t-1}$, each superset corresponding sequentially to the leaf nodes of the previous stage. The use of clusters to generate subsequent supersets, ensures that there is a stage-wise link between the nodes of the generated multistage scenario tree. Every parent node at stage t in the scenario tree branches out to leaf nodes at stage $t+1$; these leaf nodes are obtained through the reduction of supersets containing points from the same profiles for which scenario reduction was previously performed to give the parent node at stage t.

```
Algorithm 3: Optimal scenario tree generation using Sinkhorn distance
Input: Scenario fan $X_{q,\tau}$ where $q$: profile number, $\tau$: time step
       Number of stages $t = 1, 2, ..., T$
       Branching $B_t$
Procedure:
for $t = 1 : T$ do
    Compute $n_t = n_{t-1} B_t$, where $n_0 = 1$
    for $w = 1 : n_{t-1}$ do
        if $t = 1$ then
            Define $I_w^t := X(\text{all}, t)$
        else
            Define $I_w^t := X(q', t)$ where $q'$: profile numbers of points in $C_w^{t-1}$
        end
        Using Algorithm 2, reduce $I_w^t$ to $S_w^t$, obtain clusters $C_1^t, C_2^t, ..., C_{n_t}^t$ and
        the corresponding probabilities
    end
end
Output: Scenario tree with stage-wise leaf nodes $S_w^t$, and corresponding
 probabilities $Pr(S_w^t)$
```

Consider a simple 3 stage example as shown in Figure 2.5. The branching structure in this tree is $\{2, 3, 2\}$ for stages 0, 1 and 2, respectively.

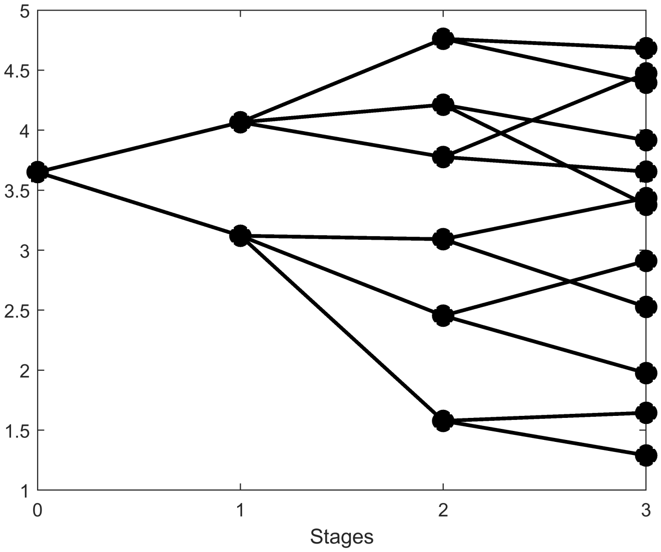


Figure 2.5: An example scenario tree

At stage 1, the superset comprises values from all profiles of the scenario fan at the time step corresponding to stage 1. This superset is to be reduced to $B_1 = 2$ points using Algorithm 2. Therefore, the optimal subset of nodes at stage 1 contains 2 points, and therefore, 2 clusters, where each point occurs is associated with a probability of occurrence equal to the ratio of number of points in the corresponding cluster and in the superset. At stage 2, two different supersets must be reduced to $B_{t=2} = 3$ points each. These supersets comprise values from those profiles clustered around each node at the previous stage $t = 1$, at the time step corresponding to stage 2. Similarly, at stage 3, six different supersets corresponding to each of the leaf nodes at stage 2, must be reduced to $B_{t=3} = 2$ points each.

Illustrating example

Random parameter occurrences can be described by a scenario fan simulation over time, as shown in Figure 2.6. Each strand of the scenario fan depicts a random occurrence of the parameter over the entire time horizon. This scenario fan can be assimilated into a simpler scenario tree structure.

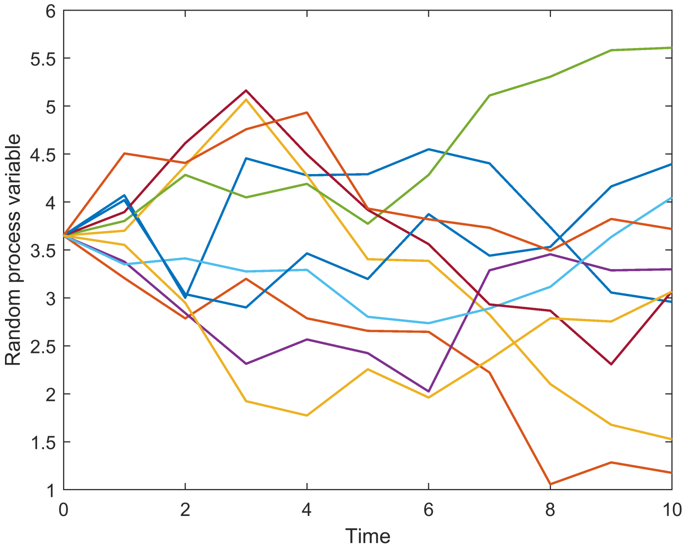
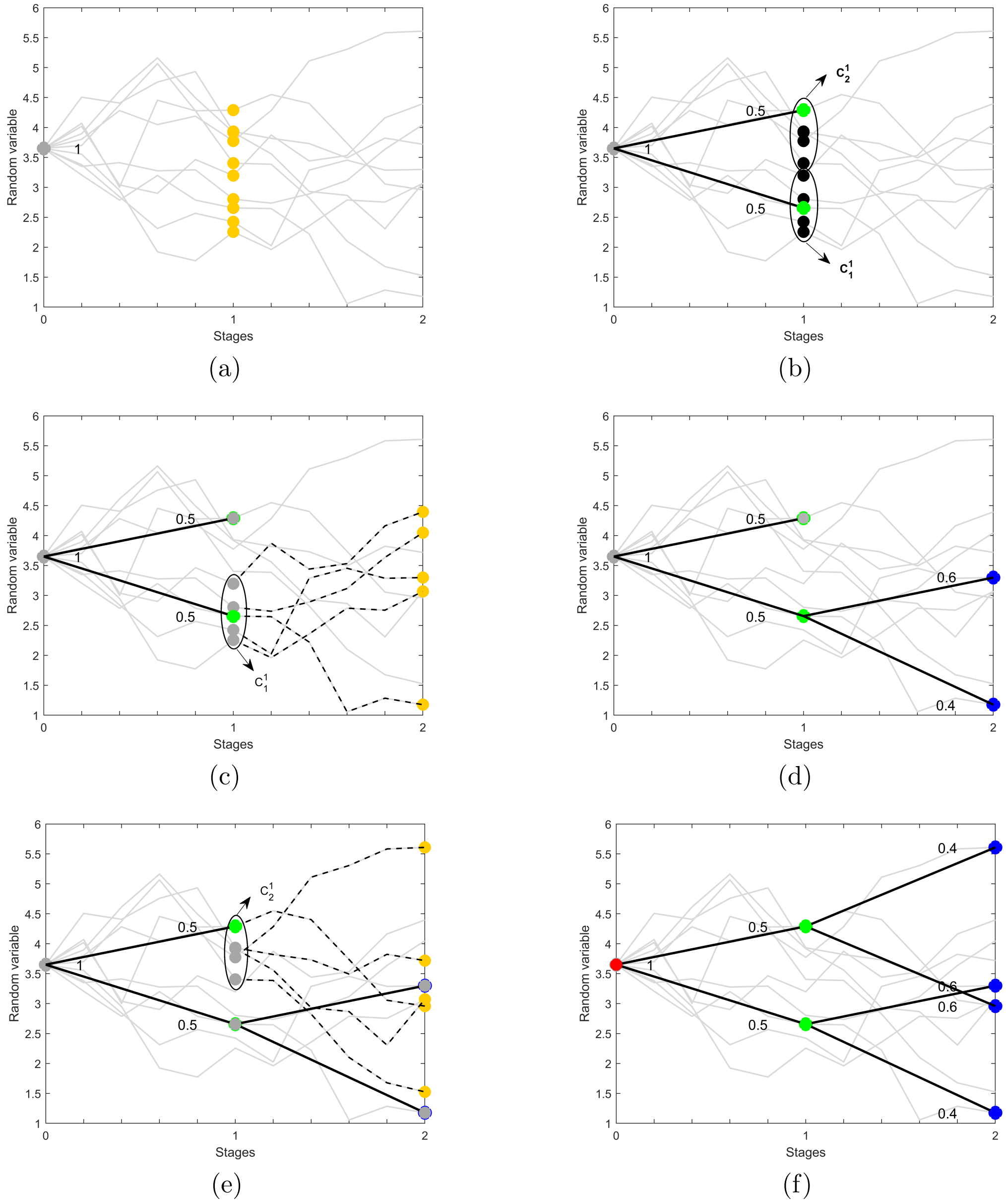


Figure 2.6: Scenario fan generation

The scenario fan in Figure 2.6 contains 10 profiles, and is generated using a base value of 3.65 with a random walk model, as follows,

$$x_{t+1} = x_t + 0.25e \tag{2.15}$$

Here, e is a random number generated from the standard Gaussian distribution.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 2.7: Multistage scenario tree generation

In this example, the scenario fan $X_{q,\tau}$ is a matrix containing 10 rows, and 11 columns spanning time steps 0 to 10. The task is to construct a 2-stage scenario tree from this fan, where each stage occurs after 5 time steps; i.e., columns 6 and 11 correspond to stages 1 and 2. It must be noted that the intermediate values or information between stages is not taken into account while generating the scenario tree, and only values at the time steps (τ) corresponding to the stages (t) are used. At each stage, we consider the branching to be $B_t = 2$.

Using the algorithm presented in Section 2.5, the multistage scenario tree (Figure 2.7) is obtained as follows: we start with stage $t = 1$. The number of nodes at stage $t = 1$ is given by $n_1 = n_0 B_1$. The number of nodes at stage 0 (n_0) is taken to be 1, by default, and this node occurs with a probability of 1. Therefore, $n_1 = 1 \times 2 = 2$. In order to obtain the nodes at stage 1 of the scenario tree, it is necessary to assess the number of supersets to be reduced at stage 1; at any stage t there are as many supersets to reduce as there are nodes in the previous stage n_{t-1}. Therefore, at stage 1, there is $n_0 = 1$ superset to reduce, and the inner loop variable, w, spans just 1 iteration. Each superset is denoted by I_w^t where w denotes the superset index corresponding to each node at the previous stage, sequentially, and t denotes the stage. At stage 1, since there is only one superset, I_1^1 comprises the values from all rows of X whose column pertains to stage 1, i.e., $X_{[1,2,\ldots,10],6}$ (yellow points in Figure 2.7a). The points in the superset I_1^1 are reduced to a subset containing $B_1 = 2$ points using Algorithm 2, which gives the optimal subset containing the points {2.6560, 4.2895}, with probabilities [0.5,0.5], as well as clusters C_1^1 and C_2^1 formed around the two points (Figure 2.7b). These clusters are formed based on the maximum probability transfer links in the optimal transport plan. Each cluster contains the profile numbers q' as well the values of the points at t. Here, the point 2.6560 is the cluster center for C_1^1 containing profile numbers (2,4,6,8,10) with corresponding values at $t = 1$ as {2.6560,2.4237,2.8028,3.1973,2.2557}, and the point 4.2895 for C_2^1 containing profile numbers (1,3,5,7,9), with values {4.2895,3.4032,3.7732,3.9175,3.9316}, respectively. The profile numbers in these clusters determine the profiles comprising each superset in the following stage. The corresponding links between parent and leaf nodes at each stage are made, and the probabilities associated with the leaf nodes are recorded from Algorithm 2.

At stage $t = 2$, there are $n_{t-1} = n_1 = 2$ supersets to be reduced, each corresponding to the nodes {2.6560,4.2895} in stage 1, respectively. Therefore, w spans the range $[1, 2]$, and the supersets to be reduced are I_1^2 and I_2^2, respectively. I_1^2 corresponds to the first node, 2.6560, of the previous stage $t = 1$ (yellow points in Figure 2.7c); therefore, I_1^2 comprises the values from that column of X pertaining to stage $t = 2$, whose row numbers are the profile numbers corresponding to the points in the first cluster C_1^1, i.e., $X_{[2,4,6,8,10],11}$. It is reduced to the optimal subset containing the points {3.2895,1.1758} (Figure 2.7d). Similarly, I_2^2 corresponds to the second node, 4.2895, which is the cluster center of C_2^1), and contains the points $X_{[1,3,5,7,9],11}$ (Figure 2.7e). It is reduced to the optimal subset {2.9582,5.6080}. The total number of leaf nodes at the final stage determines the number of profiles in the generated scenario tree (Figure 2.7f); in this case, the generated scenario tree has 4 leaf nodes in the final stage $t = 2$.

2.5.2 Process distance for quality evaluation

Consider two probability distributions, $p \in P_1$ and $p' \in P_2$. Let the support sets for the distributions be Σ_1 and Σ_2, both defined on vector space Ξ. Then, the Kantorovich Distance (KD) between the distributions is the optimal objective function value of the optimization problem presented in Model 2.4.

This does not take into account the amount of information being revealed at each stage, with respect to the filtration.

The Kantorovich Distance (KD) can be extended to stochastic processes. Consider a stochastic process in finite time, ξ_t, where $t = 0, 1, 2..., T$. The information available at time t is denoted by $v_t = (\eta_0, \eta_1, ..., \eta_t)$. Here, ξ_t can actually be denoted as a function of v_t. The sigma algebra generated by v_t is denoted by F_t. The sequence of increasing sigma algebras is known as a filtration, $F = (F_t)_{t=0}^T$.

The multistage distance, or the Process Distance (PD) is a more suitable metric to judge the similarity of two filtered probability measures, $p \in P_1$ and $p' \in P_2$, and is the optimal objective function value of the optimization problem ((Pflug and Pichler 2012),(Beltrán *et al.* 2017)),

$$\begin{aligned}
PD = \min_{\pi_{i,j}} \quad & \sum_{i \in N_T, j \in N'_T} c_{i,j}\pi_{i,j} && && (2.16a)\\
\text{s.t.} \quad & \sum_{j \in N'_T : f \to j} \pi_{i,j} = \frac{p_i}{p_e} \sum_{i' \in N_T : e \to i'} \sum_{j' \in N'_T : f \to j'} \pi_{i',j'} && \forall p_i \in P_1, (e \to i, f) && (2.16b)\\
& \sum_{i \in N_T : e \to i} \pi_{i,j} = \frac{p'_j}{p'_f} \sum_{i' \in N_T : e \to i'} \sum_{j' \in N'_T : f \to j'} \pi_{i',j'} && \forall p'_j \in P_2, (f \to j, e) && (2.16c)\\
& \sum_{i,j} \pi_{i,j} = 1 && && (2.16d)\\
& \pi_{i,j} \geq 0 && \forall i, j && (2.16e)
\end{aligned}$$

In Model 2.16, N_T and N'_T denote the set of all nodes at a given stage in the two scenario trees under consideration. The notations $f \to j$ and $e \to i$ denote that intermediate nodes f and e are predecessors of leaf nodes j and i. p_i, p'_j, p_e and p'_f are the probabilities of reaching nodes i, j, e, and f, respectively. The constraints (Equations 2.16b and 2.16c) impose the probability transfer between nodes of the scenario trees under comparison, at their corresponding time stages, subject to the satisfaction of conditional probability, by taking into account the sigma algebras involved.

The optimization problem in Model 2.16 accounts for probability mass transfer between corresponding time stages. This LP problem may become rather large, based on the number of stages, and the number of profiles in the original scenario fan, as well as the number of profiles in the reduced tree. It is solved in a decomposed manner, moving from one stage to the other.

Computational study

Figure 2.8 depicts the performance of the Sinkhorn distance-based scenario reduction in contrast with the K-means clustering-based reduction (Xu *et al.* 2012) available in literature, using the process distance (PD) metric. In their work (Xu *et al.* 2012), the authors constructed a multistage scenario tree from a scenario fan by performing K-means clustering of the supersets at every stage t in order to obtain the reduced nodes. The reduced subset at every iteration is updated as the mean of the points in the K-means clusters. In the comparison between the Sinkhorn distance-based approach and the K-means based approach, the scenario fan width was varied between 40 and 250 profiles, and a 4-stage scenario tree was generated in every test case, containing 32 profiles (with $B_t = 2, t = 0, 1, ..., 4$). In order to generate these profiles, for every fan width, 100 test scenario fans were generated, each of which was converted into multistage scenario trees using the two methods, and their average approximation performance was plotted in the figure. It is observed that, overall, the proposed Sinkhorn distance-based scenario reduction algorithm gives better performance that the K-means clustering-based algorithm.

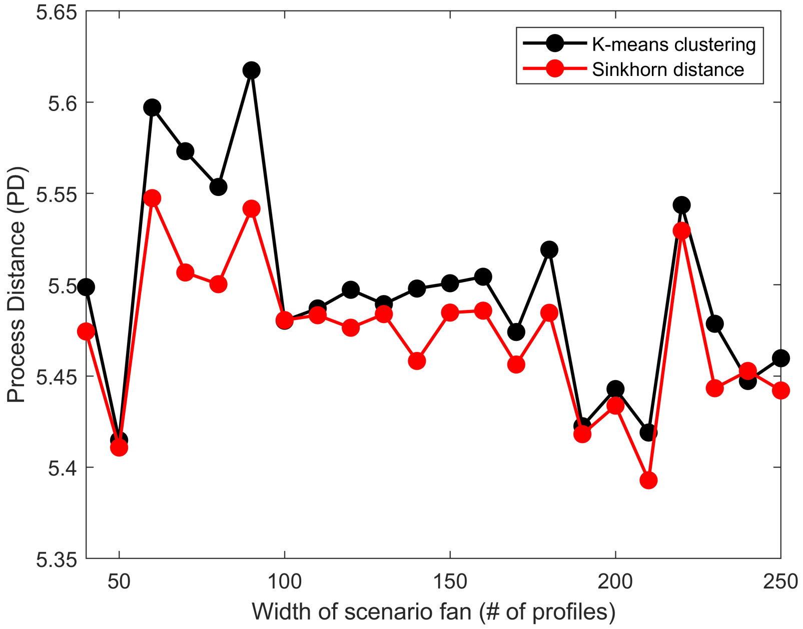


Figure 2.8: A comparison of scenario tree generation approaches using process/multistage distance (PD) for a 5-(time) stage scenario tree

2.6 Case studies

In this section, we present two case studies to demonstrate the application of the proposed optimal scenario reduction and scenario tree generation approaches. The first case study is a two-stage stochastic programming-based planning problem where the scenario reduction approach from Section 2.4.2 is applied. The second case study is a multistage stochastic programming-based chemical plant design problem where the scenario tree generation method from Section 2.5.1 is applied.

2.6.1 The farmer's problem

The farmer's problem (Birge and Louveaux 2011) presents a bench-mark case study for stochastic programming in literature; in this work, we illustrate the performance of the proposed Sinkhorn distance-based scenario reduction algorithm developed on this study. The problem deals with the allotment of 500 acres of land as farming area to be distributed between a choice of three crop species - wheat, corn, and sugar beet - under uncertainty in crop yield (y_1, y_2, y_3). To this end, x_1, x_2, and x_3 are defined as the amounts of land (in acres) allotted to wheat, corn, and sugar beet, respectively. The planting costs per acre are given as \$150, \$230, and \$260, respectively. Furthermore, constraints are placed on the problem as follows,

- Since wheat and corn are also used as cattle feed, the minimum required yield of wheat and corn are 200 tons and 240 tons, respectively.
- Wheat may be purchased at a price of \$238 per ton, while corn may be purchased for \$210 per ton. In this problem, b_1 and b_2 are defined as the amount of wheat and corn purchased (in tons), respectively.
- Wheat and corn may be sold at \$170 and \$150 per ton, respectively. Sugar beet may be sold at a price of \$36 per ton for amounts under 6000 tons, and at a lower price of \$10 per ton thereafter. In this problem, s_1 and s_2 are defined as the amount of wheat and corn sold (in tons), respectively. The amount of sugar beet sold at the higher price is defined as s_3, whereas that sold at the lower price is defined as s_4.

The optimization model for the farmer's problem may be written as (Model 2.17):

$$\begin{aligned}
\min\ & 150x_1 + 230x_2 + 260x_3 + 238b_1 - 170s_1 + 210b_2 - 150s_2 - 36s_3 - 10s_4 && \text{(2.17a)}\\
\text{s.t. } & x_1 + x_2 + x_3 \le 500 && \text{(2.17b)}\\
& y_1x_1 + b_1 - s_1 \ge 200 && \text{(2.17c)}\\
& y_2x_2 + b_2 - s_2 \ge 240 && \text{(2.17d)}\\
& s_3 + s_4 \le y_3x_3 && \text{(2.17e)}\\
& s_3 \le 6000 && \text{(2.17f)}\\
& x_1, x_2, x_3, b_1, b_2, s_1, s_2, s_3, s_4 \ge 0 && \text{(2.17g)}
\end{aligned}$$

In the scenario-based stochastic programming formulation of the problem, a finite set (K) of scenarios of crop yield ($y_{1,k}, y_{2,k}, y_{3,k} \quad \forall k \in K$) are considered. These scenarios are assumed to be based around the average yield of 2.5, 3, and 20 for wheat, corn, and sugar beet, respectively. To this end, 1000 representative scenarios are generated, as shown in Figure 2.9a. The scenario-based stochastic programming formulation of the farmer's problem is given in Model 2.18 as,

$$\min \ 150x_1 + 230x_2 + 260x_3 + \sum_{k \in K} p_k \big(238b_{1,k} - 170s_{1,k} + 210b_{2,k} - 150s_{2,k} - 36s_{3,k} - 10s_{4,k} \big) \tag{2.18a}$$

$$\text{s.t. } x_1 + x_2 + x_3 \leq 500 \tag{2.18b}$$

$$y_{1,k}x_1 + b_{1,k} - s_{1,k} \geq 200, \quad \forall k \in K \tag{2.18c}$$

$$y_{2,k}x_2 + b_{2,k} - s_{2,k} \geq 240, \quad \forall k \in K \tag{2.18d}$$

$$s_{3,k} + s_{4,k} \leq y_{3,k}x_3, \quad \forall k \in K \tag{2.18e}$$

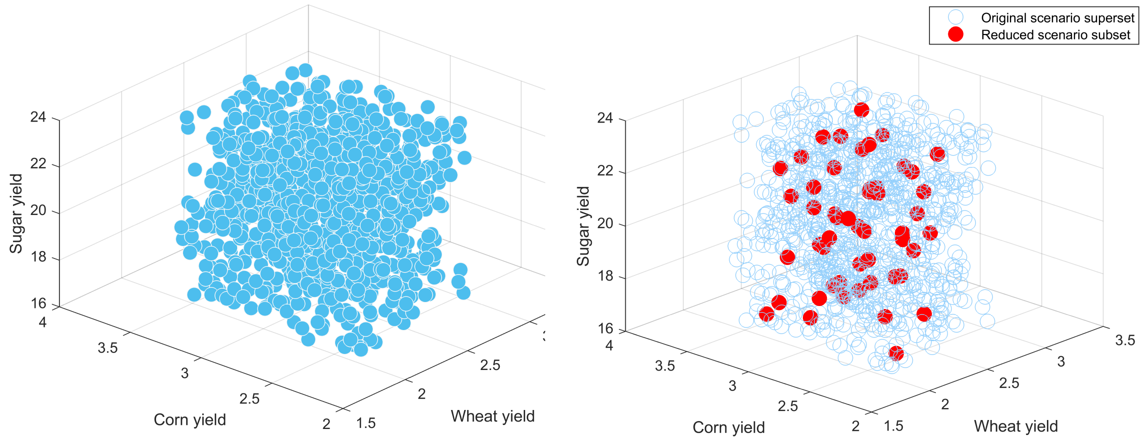
$$s_{3,k} \leq 6000, \quad \forall k \in K \tag{2.18f}$$

$$x_1, x_2, x_3 \geq 0 \tag{2.18g}$$

$$b_{1,k}, b_{2,k}, s_{1,k}, s_{2,k}, s_{3,k}, s_{4,k} \geq 0, \quad \forall k \in K \tag{2.18h}$$

Here, p_k refers to the probability of occurrence of scenario k. The probability values (p_k) are computed by the Sinkhorn distance-based scenario reduction algorithm as the probability mass allocated to each element in the set of reduced scenarios from the superset of elements.

The stochastic optimization model using the scenarios generated from the proposed Sinkhorn distance-based method was solved using the PySP (Watson *et al.* 2012) module in Python, for 30 different runs, for the same original superset of 1000 scenarios. The solutions from each of these stochastic optimization runs was extracted, and the optimal objective values were calculated.



(a) (b)

Figure 2.9: Scenario representation of 3-dimensional uncertain yield (a) 1000 scenario superset, (b) 50 scenario subset from the proposed Sinkhorn distance-based scenario reduction method (Algorithm 2)

The quality of the generated scenario set is judged by two tests: in-sample stability and out-of-sample stability (Kaut and Stein 2003). It is important to note here, that how good a scenario generation method is does not depend on how well-approximated the scenario tree is with respect to the original continuous scenario distribution; rather, it depends on how good the solution of the model using the scenario tree is, with respect to the solution using the original model.

The in-sample stability test for a scenario generation method requires that, "for a number of scenario trees generated by the chosen scenario generation method, whichever scenario tree is chosen, the optimal objective value reported across these multiple scenario trees is approximately the same". In-sample stability of a scenario generation method ensures that each time the method is used to generate a set of scenarios, it always performs to give similar optimal objective values. In-sample stability does not give any insight into how good the stochastic approximation of the model itself is, with respect to the true model; rather, it is only concerned with achieving a similar level of good each time it is used for scenario generation for a particular problem statement.

In contrast to the in-sample stability test, which compares the performance of the scenario generation method across different iterations of the method itself, the out-of-sample stability test compares the performance of the scenario generation method to the performance of the actual model itself. The out-of-sample stability test for a scenario generation method requires that, “for a number of scenario sets (trees) generated by the chosen generation method, whichever scenario tree is chosen, the optimal solution reported by each of these scenario sets is approximately equal to the true solution itself, and by extension, the optimal objective value reported by each of these multiple scenario sets is approximately equal to the true optimal objective value”. In contrast to in-sample stability, out-of-sample stability does give insight into how good the stochastic approximation of the model itself is, with respect to the true model - it gives insight into how good the scenario generation method is for a particular problem statement.

The in-sample, and out-of sample stability results for this case study are depicted in Figures 2.10a and 2.10b, respectively. Both stability tests were conducted for a number of reduced scenario subset sizes of 50, 100, 150, 200, 250, and 500 scenarios, for the same reference scenario superset of 1000 scenarios. From Figure 2.10a, it is observed that as the number of representative scenarios in the reduced subset increases, the variance in the optimal objective value decreases overall. From Figure 2.10b, it is observed that when the true problem is solved for first stage decisions fixed from the solutions of the approximated problems, the optimal objective on average reaches close to the true optimal objective value of 1.113×10^5.

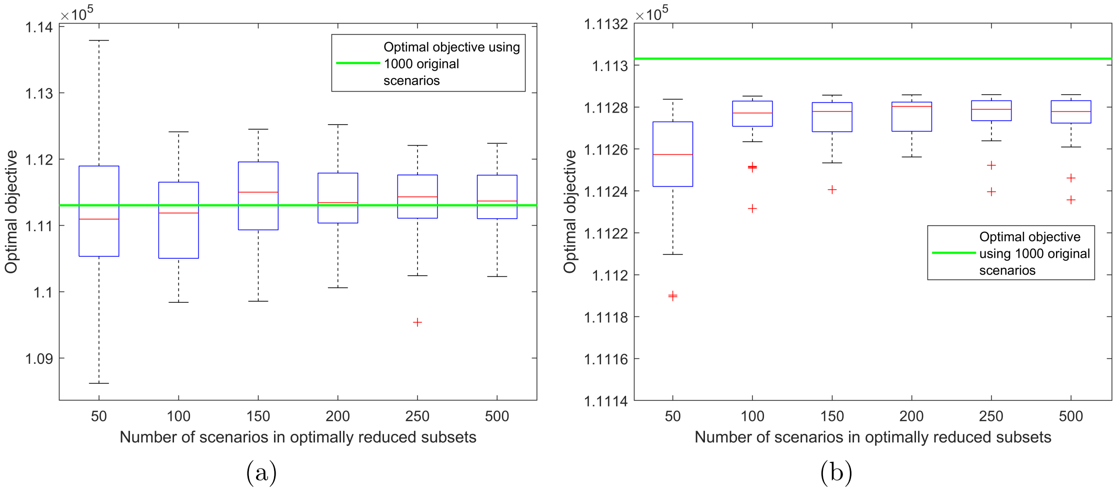


Figure 2.10: Stability test results for the 2-stage farmer's problem: (a) In-sample stability, (b) Out-of-sample stability

2.6.2 Chemical plant design problem

In this section, we illustrate the use of optimal scenario trees generated using Algorithm 3 in the multiperiod design and operation of a chemical plant under demand uncertainty, adapted from the work of Subrahmanyam *et al.* (1994).

The design superproblem in this case study is defined over the set of all tasks to be performed, denoted by $i \in I$, the set of all plant types that can be considered for construction denoted by $j \in J$, the set of all resources in play is denoted by $s \in S$, and the set of all time nodes is denoted by $t \in T$. Furthermore, I_j^{tasks} is the set of tasks that can be performed by by each plant type, while I_i^{equip} is the set of plant types that each task can be performed on.

The model contains a number of decision variables defined over the aforementioned sets. $y_{i,j,t}$ denotes the number of times task i is performed on plant type j during the time period t. Each time period is H_t days long, and each task takes $p_{i,j}$ days to complete. The number of plants of type j that come online in time period t is denoted by $n_{j,t}$, while the total number of plants of type j that are active in the time period

t is denoted by $N_{j,t}$. The cost incurred in performing tasks is denoted by $c^o_{i,j,t}$, and the total operational budget for each time period is given by C^o_t. The amount of each task that is performed on each plant type is arbitrarily measured in reaction units, and is denoted by $B_{i,j,t}$. The capacity of each plant to perform a task is given by $m_{i,j}$. The amount of resource s that is available in inventory at the end of time period t is denoted by $A_{s,t}$, and is estimated by a material balance, where the stoichiometric ratio is given by $f_{s,i}$. The amount of resource s that is sold in time period t is given by $q_{s,t}$ and the amount purchased by $z_{s,t}$. It is to be noted that only certain resources ($s = 4, 7$) are sold, and only some resources ($s = 1, 2$) are purchased. The limit on the amount of resources that can be stored in inventory at the end of any time period is given by $A^{\max}_{s,t}$. The maximum purchase limit on resources is given by $Z_{s,t}$.

Additionally, the amount of resources sold, $q_{s,t}$, is classified into two types, the amount sold below and equal to the demand, denoted by $q^0_{s,t}$, and the amount sold exceeding demand, given by $q^+_{s,t}$. Only the amount of resources sold below and equal to the demand contributes to the total profit. In this problem, the demand for resources is stochastic in nature, and takes the value $Q_{s,k,t}$, where $k \in K$ represents the set of discrete uncertain scenarios. The parameters used in the problem are given in Table 2.3. The demand for resources $Q_{4,k,t}$ and $Q_{7,k,t}$ across k scenarios is uniformly distributed between with an average of $[150, 200]$, respectively, and deviation of 7, across 1000 scenarios, which is further reduced optimally to smaller subset sizes. The complete stochastic optimization problem is formulated as Model 2.19.

$$\min \sum_{t=1}^{T} \mathbb{E}_k \left[\sum_{s=1}^{S} \left(v^{\text{sold}}_{s,k,t} q^0_{s,k,t} - v^{\text{buy}}_{s,t} z_{s,k,t} \right) - \sum_{j=1}^{J} \left(n_{j,k,t} C_{j,t} + \sum_{i=1}^{I} c^o_{i,j,t} y_{i,j,k,t} \right) \right] \tag{2.19a}$$

$$\text{s.t.} \quad \sum_{i \in I^{\text{tasks}}_j} p_{i,j} y_{i,j,k,t} \leq H_t N_{j,t} \qquad \forall j \in J, t \in T, k \in K \tag{2.19b}$$

$$N_{j,t} = \sum_{\tau=1}^{t} n_{j,k,\tau} \qquad \forall j \in J, t \in T, k \in K \tag{2.19c}$$

$$A_{s,k,t} = A_{s,k,t-1} + \sum_{i} \sum_{j \in I^{\text{equip}}_i} f_{s,i} B_{i,j,k,t} - q_{s,k,t} + z_{s,k,t} \forall s \in S, t \in T, k \in K \tag{2.19d}$$

$$A_{s,k,t} \leq A^{\max}_{s,t} \qquad \forall s \in S, t \in T, k \in K \tag{2.19e}$$

$$B_{i,j,k,t} \leq m_{i,j} y_{i,j,k,t} \qquad \forall i \in I, j \in J, t \in T, k \in K \tag{2.19f}$$

$$q_{s,k,t} = q^0_{s,k,t} + q^+_{s,k,t} \qquad \forall s \in S, t \in T, k \in K \tag{2.19g}$$

$$q_{s,k,t} \leq Q_{s,k,t} \qquad \forall s \in S, t \in T, k \in K \tag{2.19h}$$

$$z_{s,k,t} \leq Z_{s,t} \qquad \forall s \in S, t \in T, k \in K \tag{2.19i}$$

$$A_{s,k,t}, z_{s,k,t}, q^0_{s,k,t}, q^+_{s,k,t}, B_{i,j,k,t} \in \mathbb{R}^+ \qquad \forall i \in I, j \in J, s \in S, t \in T, k \in K \tag{2.19j}$$

$$n_{j,k,t}, y_{i,j,k,t} \in \mathbb{Z}^+ \qquad \forall i \in I, j \in J, t \in T, k \in K \tag{2.19k}$$

$$N_{j,t} \in \mathbb{Z}^+ \qquad \forall j \in J, t \in T \tag{2.19l}$$

As in the case of the farmer's problem case study, we solved this multistage stochastic programming problem using the PySP (Watson *et al.* 2012) module in Python. The PySP module generates a stochastic formulation of the optimization model, given the scenario structure, containing the following correspondence information: 1) variable-stage, 2) parent node-leaf node, 3) node-stage, 4) scenario-leaf node at final stage.

In this case study, we assume that the true distribution of the resource demands $Q_{s,t}$ may be represented by a set of 1000 scenarios. In order to evaluate the performance of the proposed algorithm for scenario tree generation, a number of scenario trees containing 50, 100, 150, 200, 250, and 500 profiles, respectively, were generated, and the scenario-based stochastic programming problem was solved. Figure 2.11 shows the in-sample stability results for the chemical plant case study. From the figure, we observe that as the number of scenarios in the reduced scenario tree increases from 50 to 500, the objective function value gets closer to the true objective function value of 2.6667×10^4. Figure 2.12 shows the evolution of process distance with tree size. As the number of approximating scenarios increases, the process distance decreases and gets closer to the PD value of a 1000 scenario tree.

Table 2.3: List of parameters for the chemical plant design problem

Parameter		Time period				Remarks
		$t = 1$	$t = 2$	$t = 3$		
$A_{s,t}^{\max}$	$s = 1$		400			
	$s = 2$					
	$s = 3$					
	$s = 4$					
	$s = 5$					
	$s = 6$					
	$s = 7$					
H_t		80	80	80		
$C_{j,t}$	$j = 1$					
	$j = 2$					
	$j = 3$					
$Z_{s,t}$	$s = 1$					
	$s = 2$					
$v_{4,t}^{\text{sold}}$		51	50	49		$v_{s,t}^{\text{sold}} = 0, \forall s = 1, 2, 3, 5, 6$
$v_{7,t}^{\text{sold}}$		70	71	68		
$v_{1,t}^{\text{buy}}$		23	24	25		$v_{s,t}^{\text{buy}} = 10^{20}, \forall s = 3, 4, 5, 6, 7$
$v_{2,t}^{\text{buy}}$		25	26	27		
Parameter		Plants				
		$j = 1$	$j = 2$	$j = 3$		
$p_{i,j}$	$i = 1$		4			
	$i = 2$					
	$i = 3$					
	$i = 4$					
$m_{i,j}$	$i = 1$	100	200	150		
	$i = 2$					
	$i = 3$					
	$i = 4$					
I_j^{tasks}		$1, 4$	$1, 4$	$2, 3$		
Parameter		Tasks				
		$i = 1$	$i = 2$	$i = 3$	$i = 4$	
I_i^{equip}		$1, 2$	3	3	$1, 2$	
$f_{s,i}$	$s = 1$	-1	0	0	0	
	$s = 2$	-1	0	-1	0	
	$s = 3$	1	-1	0	0	
	$s = 4$	0	1	0	0	
	$s = 5$	0	1	-1	0	
	$s = 6$	0	0	1	-1	
	$s = 7$	0	0	0	1	

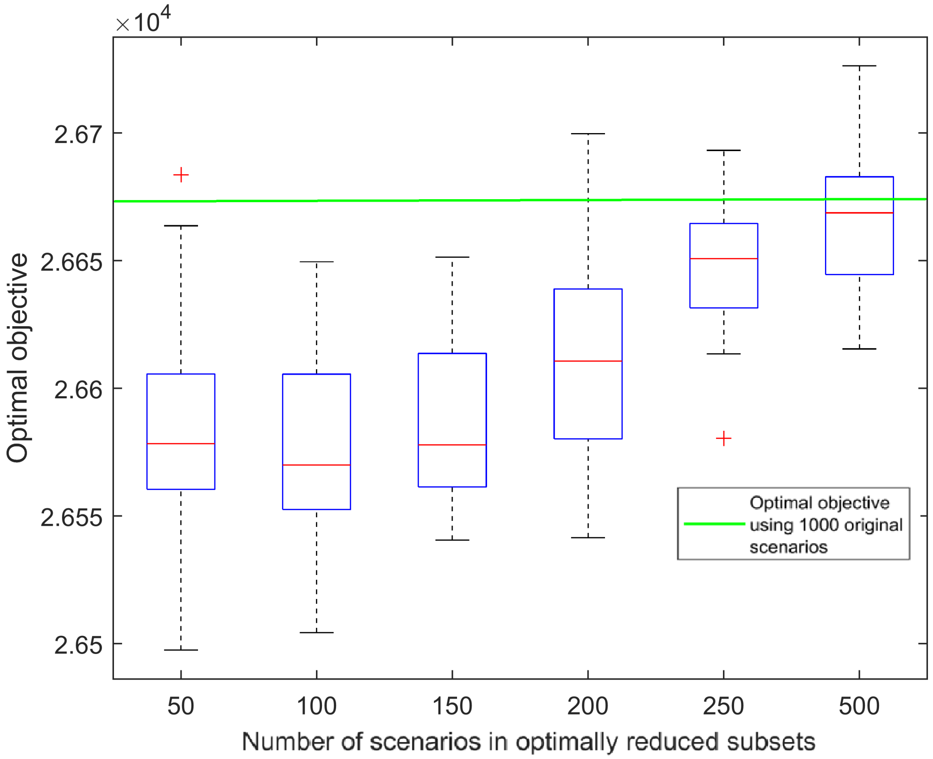


Figure 2.11: In-sample stability results for multi-period stochastic design of a chemical plant

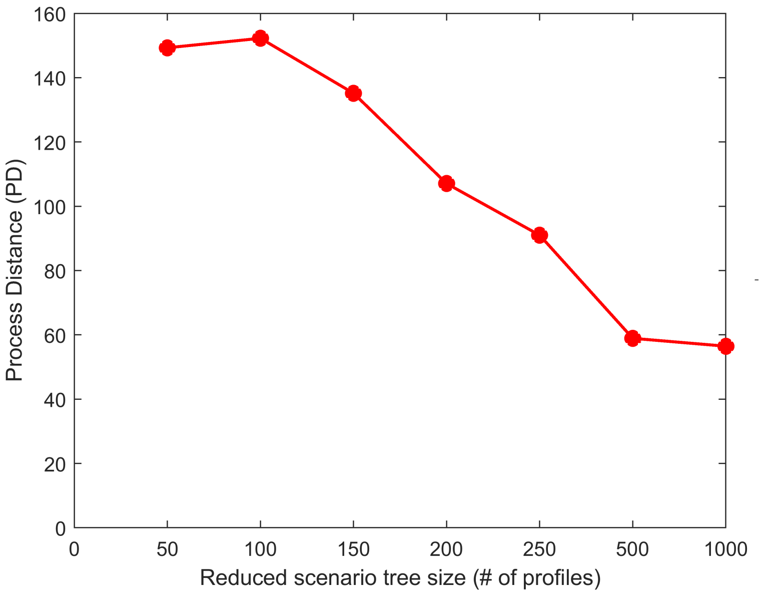


Figure 2.12: Process distance evolution with reduced scenario set size for multi-period stochastic design of a chemical plant

2.7 Conclusion

In this work, we propose a method to reduce a large set of scenarios to a smaller subset of chosen size through entropy-regularized optimal transport. We leveraged the Sinkhorn-Knopp algorithm to replace the linear programming formulation of the discrete optimal transport problem with an iterative computational procedure, to obtain the Sinkhorn distance as a quantitative measure of similarity between the probability distributions of the superset and the optimal subset. Using this measure, we developed an algorithm for optimal scenario reduction, and extended its use to scenario tree generation for multistage stochastic programming. We illustrated the proposed scenario reduction and scenario tree generation algorithms on two case studies, and assessed their stability. We found that replacing the Kantorovich Distance with the Sinkhorn distance in scenario reduction resulted in a significant decrease in computational time. Furthermore, in stochastic programming applications, the solutions obtained from using Sinkhorn distance-based scenario reduction converged to the solutions of the problems solved using the original scenario set with good accuracy. The proposed algorithms for scenario reduction and scenario tree generation are effective for scenario-based stochastic programming.

The scenario reduction, and scenario tree generation methods proposed and discussed in this chapter leveraged the “distance” or similarity metric property of (entropy-regularized) optimal transport. To this end, we treated the destination probability distribution’s support set as the decision variable to be optimized (in an iterative fashion). In Chapters 3 and 4, we pivot to a different applicability of the optimal transport distance wherein we no longer seek the explicit optimal support set of the destination distribution; rather, a variant of the optimal transport problem in incorporated into a higher-level optimization problem wherein the objective is to obtain the optimal solution to this higher-level problem subject to a “constrained” optimal transport problem. Here, the "constraint" refers to the radius of the ambiguity set, which is discussed in detail in Chapters 3 and 4.

Chapter 3

Distributionally Robust Optimization using Optimal Transport for Gaussian Mixture Models

Abstract: Distributionally robust optimization (DRO) is an increasingly popular approach for optimization under uncertainty when the probability distribution of the uncertain parameter is unknown. Well-explored DRO approaches in literature, such as Wasserstein DRO, do not make any specific assumptions on the nature of the candidate distributions considered in the ambiguity set. However, in many practical applications, the uncertain parameter may be sourced from a distribution that can be well modeled as a Gaussian Mixture Model (GMM) whose components represent the different subpopulations the uncertain parameter may belong to. In this work, we propose a new DRO method based on an ambiguity set constructed around a GMM. The proposed DRO approach is illustrated on a numerical example as well as a portfolio optimization case study for uncertainty sourced from various distributions. The results obtained from the proposed DRO approach are compared with those from Wasserstein DRO, and are shown to be superior in quality with respect to out-of-sample performance.

3.1 Introduction

Mathematical optimization refers to the task of finding the best set of decisions that minimize a defined cost pertaining to the system, over a set of constraints that describe systems including but not limited to physical or chemical systems, manufacturing processes and supply chains. In a practical setting, however, most systems are affected by uncertainty. Therefore, the optimization must take this perturbation into consideration. Optimization under uncertainty is of immense importance for real-world applications and therefore, a considerable amount of research has been devoted to this task (Sahinidis 2004; Ning and You 2019; Keith and Ahner 2021).

Optimization problems under uncertainty are mainly tackled through the lens of whether the probability distribution of the underlying uncertainty in a system is available. If the probability distribution of the uncertainty is readily available or can be reasonably estimated from sampled data, the problem may be tackled through stochastic programming (Dantzig 1955). When the support of the uncertainty can be considered a finite and discrete set of realizations, this strategy is termed scenario-based stochastic programming. Significant strides have been made in optimization literature regarding this technique which provides a framework to incorporate sampled information about the underlying uncertainty into the optimization process (Wallace and Ziemba 2005). However, the quality of this approach relies heavily on the approximation of the underlying probability distribution through the set of scenarios (Esfahani and Kuhn 2018; Shapiro and Nemirovski 2005). While approaches such as sample average approximation (SAA) offer certain finite sample performance guarantees under certain assumptions (Shapiro and Nemirovski 2005), the out-of-sample performance of such approximations proves lacking in situations where the sample size is small. In contrast, another common way to tackle optimization under uncertainty is through robust optimization (Ben-Tal and Nemirovski 1998). In this approach, no information about the underlying probability distribution of the uncertainty is used; rather, the problem hedges against the worst-case realization of sampled uncertainty. While this approach eliminates the need for a probability distribution, it requires efficient design of the uncertainty set to avoid overly conservative solutions.

Distributionally robust optimization (DRO) was developed as an intermediate approach to optimization under uncertainty that combines facets of both stochastic programming as well as robust optimization to combat the drawbacks of each technique. This approach was first utilized by Scarf 1958 to solve the newsvendor optimization problem for worst-case profit maximization. Since his seminal work on DRO, a significant amount of research has been done on this topic (Rahimian and Mehrotra 2019). The distinguishing feature of distributionally robust optimization (DRO), also referred to as 'ambiguous stochastic optimization' in literature, is the treatment of the underlying distribution of the uncertainty in the problem. Unlike in stochastic programming wherein the estimated distribution is assumed to be a reasonable approximation, DRO introduces a level of ambiguity into this step by considering an 'ambiguity set' (Wiesemann *et al.* 2014) of distributions centered on this approximation. Therefore, DRO bypasses the approximation drawbacks of stochastic programming, while still utilizing some amount of probabilistic information about uncertainty, unlike in robust optimization which discards it entirely, to give better optimal solutions. More specifically, DRO aims to hedge against the worst-case expectation of the objective function, further denoted as the loss function, over the ambiguity set of probability distributions (termed as candidate distributions) that is constructed using the approximated 'nominal' distribution supported on sampled uncertainty (Wiesemann *et al.* 2014). Furthermore, for convex and compact ambiguity sets and real-valued loss functions, the worst-case expectation in DRO is found to be equivalent to a coherent risk measure (Artzner *et al.* 1999), thus establishing the connection between DRO and risk averse optimization (Ruszczyński and Shapiro

2006). When DRO is considered in a chance-constrained setting, it refers to a distributionally robust risk-averse optimization approach in which a small probability of violation of constraints affected by uncertainty is allowed so as to avoid infeasibility due to hard violations. In this approach, the allowed probability of violation of constraints is enforced over the ambiguity set accounting for distributional uncertainty. To this end, distributionally robust chance-constrained programming (DRCCP) is a topic that has garnered significant interest in recent times (Hota *et al.* 2019; Yang and Li 2022; Chen *et al.* 2022).

The performance of a DRO model is heavily governed by the choice of ambiguity set made by the modeler. In order to obtain an optimal solution that delivers good out-of-sample performance, it is necessary to construct an ambiguity set that is large enough so as to contain the true underlying distribution with a good level of certainty but not so large as to consider pathological distributions that contribute to overly conservative solutions (Esfahani and Kuhn 2018). The construction of an ambiguity set for DRO may be moment-based, shape-preserving, kernel-based, or metric-based (Rahimian and Mehrotra 2019). Moment-based ambiguity sets consider those probability distributions whose statistical moments satisfy certain properties (Ghaoui *et al.* 2003; Goldfarb and Iyengar 2003; Grunwald and Dawid 2004; Delage and Ye 2010; Goh and Sim 2010; Natarajan and Teo 2017). When the ambiguity set is constructed to be shape-preserving, it includes all candidate distributions that have similar structural properties to those of the nominal distribution (Popescu 2005; Parys *et al.* 2016). Kernel-based ambiguity set construction seeks to build an ambiguity set containing all candidate distributions formed through a kernel, whose parameters are similar to those of the nominal distribution (Bertsimas and Kallus 2020; Zhu *et al.* 2021). Metric-based ambiguity sets - using similarity measures such as optimal transport distances, maximum mean discrepancies, l_p norm-based distances, and ϕ divergences, and total variation distances - treat the nominal distribution to be the ‘center’ around which a neighborhood of candidate distributions that are a certain $\epsilon-$level of similarity to the center are considered (Pflug and Wozabal 2007; Hanasusanto and Kuhn 2013; Bayraksan and Love 2015; Abadeh *et al.* 2015; Esfahani and Kuhn 2018; Jiang and Guan 2018; Gao and Kleywegt 2022; Blanchet *et al.* 2022). This ϵ is a hyperparameter that is defined by the user, and is an indicator of how ‘ambiguous’ we consider the approximated distribution to be.

Ambiguity set construction via the optimal transport distance has been studied extensively over the past decade (Pflug and Wozabal 2007; Mehrotra and Zhang 2014; Esfahani and Kuhn 2018; Gao and Kleywegt 2022; Chen *et al.* 2022). Esfahani and Kuhn 2018 demonstrated the finite convex reformulation available for 1-Wasserstein distance-based DRO ambiguity sets, and presented interesting out-of-sample results. More recently, Li and Mao 2022 studied the different choices of p^{th} order Wasserstein metrics available for DRO, and introduced a new class of coherent Wasserstein metrics for DRO. Liu *et al.* 2022 used Wasserstein DRO in the context of power scheduling under pricing as well as wind power uncertainties, wherein the authors show that the DRO method outperforms stochastic optimization in terms of its out-of-sample performance.

In many applications, the underlying data distribution may not be fully known or may be subject to variability, and accurately modeling this uncertainty is crucial for achieving good performance in DRO. Gaussian Mixture Models (GMMs) provide a flexible and powerful framework for modeling uncertain data distributions, as they are known to capture complex and multimodal distributions with relatively few parameters. Using GMMs to model uncertainty in DRO may thus help in improving the robustness and generalizability of optimization algorithms, by allowing them to effectively handle uncertain or variable data distributions. The objective of this work is to explore the usage of GMMs in modeling ambiguity sets for DRO applications.

Contributions of the presented work include:

- A novel approach for distributionally robust optimization wherein the ambiguity set is constructed using a variant of optimal transport distance between Gaussian Mixture Models (Chen *et al.* 2018).

- A tractable formulation of the proposed DRO problem, henceforth referred to as the W_d-DRO problem, under the OT-GMM-based ambiguity set and a numerical example illustrating its use.

- Comparison of the proposed OT-GMM-based DRO method to the established Wasserstein DRO over a portfolio optimization case study to illustrate the improvement in out-of-sample performance.

- Results on the worst-case expectation distribution in the proposed DRO method and a simple method to compute an upper bound on the radius of the ambiguity set, an important hyperparameter in DRO that affects the conservativeness of the solution.

The rest of the chapter is organized as follows; we introduce the relevant theory in Section 3.2 and the derivation of the tractable formulation of the proposed W_d-DRO method in Section 3.3. In Section 3.4, we present a numerical illustrative study which we use to discuss the worst-case expectation distribution and a protocol for the upper bounding radius of the W_d-DRO ambiguity set. Section 3.5 compares and contrasts the performance of the established Wasserstein DRO and the proposed W_d-DRO methods for a portfolio optimization case study in the context of in-sample and out-of-sample performance. Finally, in Section 3.6, we present a summary of the work undertaken in this chapter.

3.2 Preliminaries

3.2.1 Ambiguity set and DRO

A metric-based ambiguity set may be defined as a ball in the space of probability distributions by specifying a center and a radius. The center of the ball represents the most plausible, or the prior belief about the probability distribution, while the radius represents the degree of uncertainty or ambiguity surrounding this belief. The ambiguity set is shown schematically in Figure 3.1.

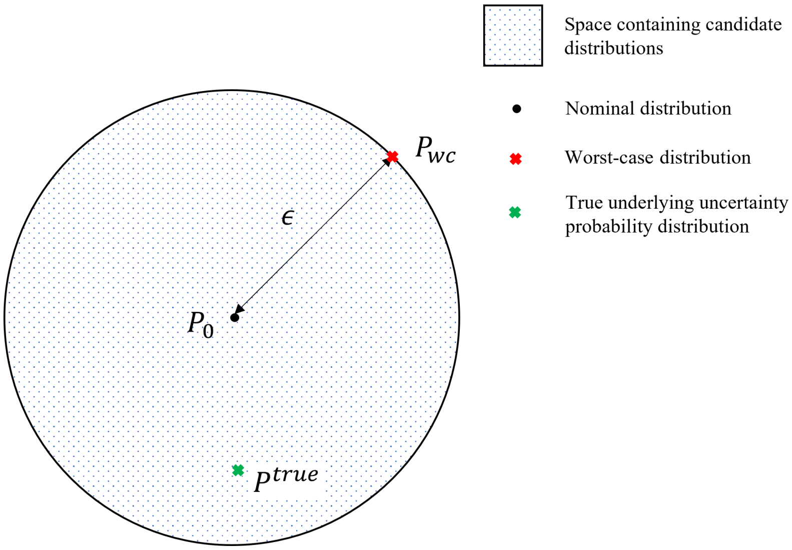

▨ Space containing candidate distributions
● Nominal distribution
✖ Worst-case distribution
✖ True underlying uncertainty probability distribution
P_{wc}
ϵ
P_0
P^{true}

Figure 3.1: An illustration of the ambiguity set and the various probability distributions of interest in the distributionally robust optimization (DRO) framework

With a defined ambiguity set of uncertainty, denoted by $\mathcal{P}$, the general distributionally robust optimization (DRO) problem may be given by

$$\min_{x\in X} \quad \max_{\mathbb{P}(\xi)\in\mathcal{P}} \quad \mathbb{E}_{\mathbb{P}(\xi)}\Big[f(x,\xi)\Big] \tag{3.1}$$

Here, we only consider the objective function to be affected by uncertainty $\xi \in \Xi$, where Ξ is a measurable continuous support. X denotes the feasible region for the decision variables x.

3.2.2 Optimal transport

Optimal transport, as a mathematical topic, has a rich history dating back to the eighteenth century when Monge (1781) posed the problem in the context of allocation of quarried soil under minimum transportation cost. In recent years, optimal transport has regained interest in several fields such as computer vision, and statistical as well as machine learning due to progress in computational ability that is able to overcome the curse of dimensionality that the conventional formulation of the problem possesses. In particular, entropy regularization of the optimal transport problem (Cuturi 2013; Clason *et al.* 2021) has gained recognition for its ability to lower computational cost of optimal transport from $O(n^3 \log n)$ in its conventional setting to $O(n^2 \log n)$ owing to the availability of the Sinkhorn Algorithm (Sinkhorn 1967), thus enabling its usage in large scale data analysis and machine learning. A number of variants of optimal transport have also been studied recently - such as unbalanced optimal transport (Benamou 2003; Caffarelli and McCann 2010; Blondel *et al.* 2018; Chizat *et al.* 2018), semi-discrete optimal transport (Oliker and Prussner 1989; Mérigot 2011; Lévy 2015), and multi-marginal optimal transport (Pass 2012; Pass 2015; Nenna 2016; Haasler *et al.* 2021).

Arising from the optimal transport problem, the so-called p-Wasserstein distance between any two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ is given by

$$W_p(\mathbb{P}_1, \mathbb{P}_2) := \left(\inf_{\pi \in \Pi(\mathbb{P}_1, \mathbb{P}_2)} \int_{\Xi \times \Xi} ||\xi_1 - \xi_2||^p d\pi(\xi_1, \xi_2) \right)^{1/p} \tag{3.2}$$

Here, ξ_1 and ξ_2 are the uncertain parameters that belong to the probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively. That is, $\mathbb{P}_1$ and $\mathbb{P}_2$ are marginal distributions, and $\Pi(\mathbb{P}_1, \mathbb{P}_2)$ denotes the set of joint distributions π on the space $\Xi \times \Xi$. For two random variables (or vectors of the same dimension) following Gaussian distributions $\mathbb{P}_1 := \mathcal{N}(\mu_1, \Sigma_1)$ and $\mathbb{P}_2 := \mathcal{N}(\mu_2, \Sigma_2)$, the squared 2-Wasserstein distance has a closed form analytical expression as follows,

$$W_2^2(\mathbb{P}_1,\mathbb{P}_2) := \|\mu_1 - \mu_2\|_2^2 + \mathrm{Tr}\Bigg(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Bigg) \tag{3.3}$$

3.2.3 Wasserstein DRO

The p-Wasserstein distance-based ambiguity set of radius ϵ_W is defined as

$$\mathcal{P} = \{\mathbb{P} : W_p(\mathbb{P},\mathbb{P}_0) \le \epsilon_W\} \tag{3.4}$$

When the ambiguity set $\mathcal{P}$ is constructed using the 1-Wasserstein metric (W_1), the DRO problem presented in (3.1) may be rewritten as

$$\min_{x\in X} \quad \max_{\mathbb{P}(\xi)} \quad \mathbb{E}_{\mathbb{P}(\xi)}\Big[f(x,\xi)\Big] \tag{3.5a}$$

$$\text{s.t.} \quad W_1(\mathbb{P},\mathbb{P}_0) \le \epsilon_W \tag{3.5b}$$

The aim of Wasserstein DRO is to provide an optimal solution by hedging against the expectation taken over the worst-case distribution from an ambiguity set of candidate distributions $\mathcal{P}$, which is constructed as a Wasserstein ball of radius ϵ_W centered around the nominal distribution $\mathbb{P}_0$, and contains the various corresponding candidate distributions $\mathbb{P}(\xi)$. When there is no prior knowledge or assumptions about the distribution of the uncertain parameters, a common approach for setting the nominal distribution ($\mathbb{P}_0$) in distributionally robust optimization is to use empirical distributions. An empirical distribution estimated from data assigns equal probability mass to each observed sample. Estimating such an empirical distribution to further serve as a nominal distribution for DRO provides a more realistic and data-driven approach to decision making under uncertainty.

The tractable form of the Wasserstein DRO problem may be obtained by first substituting the 1-Wasserstein optimal transport problem into constraint 3.5b, and taking the dual of the inner maximization problem in Model 3.5. Model 3.6 presents the tractable form of the Wasserstein DRO problem when the loss function is an affine function in uncertainty as $f(x,\xi) \coloneqq a(x)^{\mathrm{T}}\xi + b(x)$. A detailed derivation of the tractable forms of the Wasserstein DRO model for different support sets for the underlying uncertainty can be found in Yang and Li (2022).

$$\min_{x,\eta\geq 0,z_k} \quad \eta\epsilon_W + \frac{1}{N}\sum_{k=1}^{N} z_k \tag{3.6a}$$

$$\text{s.t.} \quad z_k \geq f\big(x, \xi_k^0\big), \quad \forall 1 \leq k \leq N \tag{3.6b}$$

$$\|a\|_* \leq \eta \tag{3.6c}$$

$$x \in X \tag{3.6d}$$

In Model 3.6, the parameter $\xi_k^0, k = 1, \cdots, N$ refers to the available data samples on uncertainty. η and z are the dual variables. The dual norm $\|\cdot\|_*$ in constraint 3.6c depends on the norm used in the ground cost of the 1-Wasserstein optimal transport problem (Equation 3.2). In this work, the 2-norm cost was used and its dual norm is also 2-norm. The structure of the Wasserstein DRO model is dependent on the loss function; for an affine loss function in ξ, and linear constraints in $x \in X$, Model 3.6 is a quadratically constrained optimization problem (QCP), which may be solved using solvers such as CPLEX or XPRESS. It may be noted that setting $\epsilon_W = 0$ in Model 3.6 effectively converts the DRO model to a stochastic programming model solved using the sample average approximation (SAA) approach, as follows,

$$\min_{x,z_k} \quad \frac{1}{N}\sum_{k=1}^{N} z_k \tag{3.7a}$$

$$\text{s.t.} \quad z_k \geq f\big(x, \xi_k^0\big), \quad \forall 1 \leq k \leq N \tag{3.7b}$$

$$x \in X \tag{3.7c}$$

3.3 GMM-based DRO

3.3.1 Gaussian Mixture Models

A Gaussian mixture model refers to a parametric probability density function that is denoted as a weighted sum of a finite number of Gaussian component density functions. Due to their ease of representation as well as their capability to perform as universal estimators (Aragam *et al.* 2018), GMMs are a powerful tool notably used for clustering and pattern recognition applications. A typical Gaussian mixture model defined over a support $x \subseteq \mathbb{R}^d$, containing L components may be represented as

$$\mathbb{P}(x) := \sum_{l=1}^{L} w_i \mathbb{G}_l(x) \tag{3.8}$$

Here, each component density $\mathbb{G}_l, \quad \forall 1 \leq l \leq L$ may be represented as multivariate Gaussian function with mean vector μ_i and covariance matrix Σ_l respectively, such that the component weights w_l sum up to 1,

$$\mathbb{G}_l(x) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_l|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(x - \mu_l)' \Sigma_l^{-1} (x - \mu_l) \right\} \tag{3.9}$$

Gaussian mixture models have several variants based on the defined structure of their parameters, namely component means and variances. The covariance matrices of the Gaussian components may be assumed (by the user) to be full ranked, or may be restricted to diagonal matrices, depending on whether the data may be assumed to be correlated or uncorrelated, respectively. Furthermore, component attributes, most commonly the covariance matrix, may be shared between the L components. A GMM is most commonly estimated using the expectation-maximization (EM) algorithm. The EM algorithm is a well-established method of finding the attributes of the components in the GMM. It is an iterative two-step method: first, the maximum likelihood of a sampled data point belonging to each Gaussian density is estimated; second, the GMM attributes are optimized through a maximization problem (Dempster *et al.* 1977; Bishop and Nasrabadi 2006). Additionally, other methods such as maximum a-posteriori (MAP) parameter estimation are available to fit GMMs to data for pattern recognition applications (McLachlan *et al.* 2019).

3.3.2 W_d metric between two GMM distributions

Consider two probability distributions, $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$, modeled as Gaussian Mixture Models (GMMs) as follows

$$\mathbb{P}^{(1)} := \; w_1^{(1)}\nu_1^{(1)} + w_2^{(1)}\nu_2^{(1)} + ... + w_{L^{(1)}}^{(1)}\nu_{L^{(1)}}^{(1)} \tag{3.10a}$$

$$\mathbb{P}^{(2)} := \; w_1^{(2)}\nu_1^{(2)} + w_2^{(2)}\nu_2^{(2)} + ... + w_{L^{(2)}}^{2}\nu_{L^{(2)}}^{(2)} \tag{3.10b}$$

The notation w refers to the weights of the Gaussian components ν (of the same dimension) in each GMM, while $L^{(1)}$ and $L^{(2)}$ refer to the number of Gaussian components used to model the support of the GMMs $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$, respectively; they may also be interpreted as the number of subpopulations identified in the supporting data. The superscripts (1) and (2) in Equations 3.10a - 3.10b refer to the probability distributions, while the subscripts denote Gaussian component indices; e.g, $w_1^{(1)}$ refers to the weighting proportion of the first Gaussian component of probability distribution $\mathbb{P}^{(1)}$ and $v_1^{(1)}$ refers to the first Gaussian component of $\mathbb{P}^{(1)}$.

The Wasserstein distance between two GMM distributions $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ (Equation 3.2) is computed as the optimal transport distance between the elements of their corresponding support sets. However, when $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ may be modeled as GMMs, Chen *et al.* (2018) have shown that the optimal transport problem may be formulated as,

$$W_d^2(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}) = \delta := \min_{\pi_{l,l'}} \quad \sum_{l=1}^{L^{(1)}} \sum_{l'=1}^{L^{(2)}} c_{l,l'}\pi_{l,l'} \tag{3.11a}$$

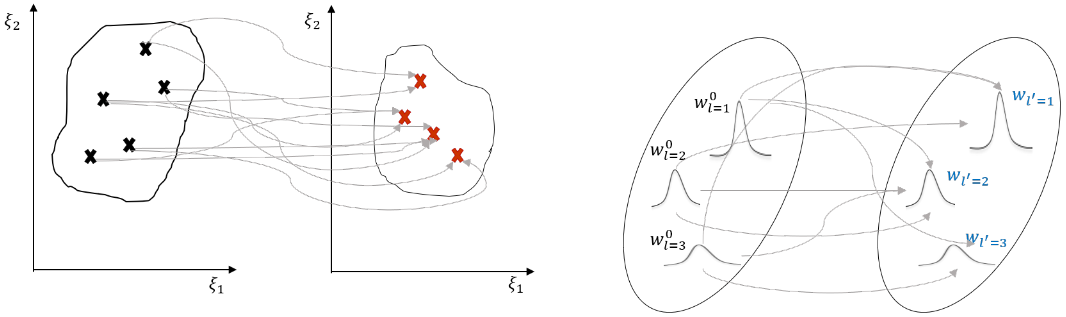
$$\text{s.t.} \quad \sum_{l'=1}^{L^{(2)}} \pi_{l,l'} = w_l^{(1)}, \quad \forall 1 \le l \le L^{(1)} \tag{3.11b}$$

$$\sum_{l=1}^{L^{(1)}} \pi_{l,l'} = w_{l'}^{(2)}, \quad \forall 1 \le l' \le L^{(2)} \tag{3.11c}$$

$$\pi_{l,l'} \ge 0, \quad \forall 1 \le l \le L^{(1)}, 1 \le l' \le L^{(2)} \tag{3.11d}$$

The W_d metric used in this work is the square root of the optimal objective (Equation 3.11a) of the OT-GMM problem (Model 3.11). The cost of transport ($c_{l,l'}$) is computed between the Gaussian components of the GMMs using the closed-form expression described in Equation 3.3. Specifically for the Gaussian components of $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$, $c_{l,l'}$ is calculated as the squared 2-Wasserstein distance between $\nu_l^{(1)} \sim \mathcal{N}(\mu_l^{(1)}, \Sigma_l^{(1)})$ and $\nu_{l'}^{(2)} \sim \mathcal{N}(\mu_{l'}^{(2)}, \Sigma_{l'}^{(2)})$ described as

$$c_{l,l'} := W_2^2(\nu_l^{(1)}, \nu_{l'}^{(2)}) := \|\mu_l^{(1)} - \mu_{l'}^{(2)}\|_2^2 + \mathrm{Tr}\left(\Sigma_l^{(1)} + \Sigma_{l'}^{(2)} - 2\left(\Sigma_l^{(1)^{1/2}} \Sigma_{l'}^{(2)} \Sigma_l^{(1)^{1/2}} \right)^{1/2} \right) \quad (3.12)$$



(a) (b)

Figure 3.2: A schematic comparison between the transport problems involved in (a) Wasserstein optimal transport, (b) and optimal transport between GMMs

Figure 3.2 illustrates the difference between the optimal transport problems using conventional Wasserstein setting (Figure 3.2a) and the GMM-based setting (Figure 3.2b). In the former case, transport is conducted between the elements supporting the probability distributions under consideration. In the case of OT-GMM, each Gaussian component in either GMM is considered akin to an element in the support set, and the weight of that component is treated as its probability mass which is conserved overall during transport. Additionally, in the OT-GMM problem, transport is conducted between spaces of Gaussian distributions, and not Euclidean spaces themselves.

In the OT-GMM problem, the conservation constraints are imposed on the component weighting proportions of the mixture models. Chen *et al.* (2018) have shown that

W_d is an upper bounding value on the 2-Wasserstein distance between the support sets of $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$. Furthermore, the authors emphasize that since the OT-GMM problem involves transport among the Gaussian components of the mixture models rather than between the supporting elements themselves, the problem scales with the number of subpopulations represented in the GMMs. This presents a computational advantage over the traditional optimal transport problem which scales with the size of the support sets involved. In addition, the W_d distance is shown to obey all properties of a metric.

3.3.3 W_d metric-based ambiguity set

In our work, the OT-GMM based W_d metric is used to construct the ambiguity set when the presence of subpopulations, or multiple modes are known or observable in the sampled uncertainty. More specifically, when the data on the uncertain parameter contains multiple modes, an ambiguity set of a certain radius ϵ_d, may be constructed around the nominal distribution $\mathbb{P}_0$ which is the GMM fitted to the sampled uncertainty, that contains all the candidate GMMs $\mathbb{P}$ such that the W_d distance between $\mathbb{P}$ and $\mathbb{P}_0$ is at most ϵ_d.

The W_d metric-based ambiguity set, parameterized by radius ϵ_d, for DRO applications is defined as follows

$$\mathcal{P} = \{\mathbb{P} : W_d(\mathbb{P}, \mathbb{P}_0) \leq \epsilon_d\} \tag{3.13}$$

It is important to note here that in this work, we consider all the candidate GMMs ($\mathbb{P}$) to be based on the same components (L) as the nominal GMM ($\mathbb{P}_0$) as,

$$\mathbb{P}_0 := \; w_1^0\nu_1^0 + w_2^0\nu_2^0 + \ldots + w_L^0\nu_L^0 \tag{3.14a}$$

$$\mathbb{P} := \; w_1\nu_1^0 + w_2\nu_2^0 + \ldots + w_L\nu_L^0 \tag{3.14b}$$

Therefore, the various candidate distributions in this proposed ambiguity set are assumed to arise as a result of flexible component weights ($w_{l'}, \quad 1 \leq l' \leq L$) on the nominal distribution's Gaussian components ($\nu_{l'}^0, \quad 1 \leq l' \leq L$). This assumption is illustrated on an arbitrary 3-component GMM in Figure 3.3; where the weights of the components of this GMM are made flexible with the only constraint imposed being $\sum_{l'=1}^{L} w_{l'} = 1$. All the candidate distributions in the ambiguity set are around the nominal GMM distribution. They are defined on the same Gaussian components as the nominal GMM but different weights on the Gaussian components can be applied. Therefore, the quality of the ambiguity set is affected by the goodness of fit of the nominal distribution. However, the method can naturally accommodate some fitting error since the candidate distribution can have various weights.

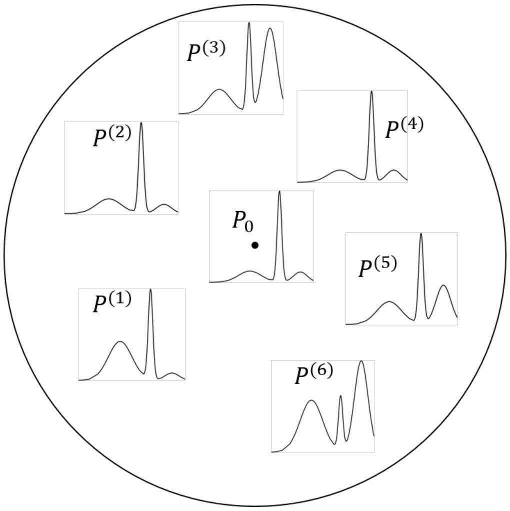


Figure 3.3: An illustration of some candidate distributions supported on the nominal GMM's ($\mathbb{P}_0$) Gaussian components with flexible weights $w_{l'}$.

3.3.4 GMM-based DRO (W_d-DRO)

In the context of DRO, You *et al.* (2021) have used GMMs to develop a data-driven ambiguity set by accounting for variable parameters, namely mean, variance, and component weights. The authors allow the aforementioned parameters to vary within their own credible regions defined by confidence thresholds to achieve a distributionally robust framework for their CVaR-based optimal power flow problem for which they further proposed a novel cutting-plane approach.

In our work, the DRO problem under the W_d metric-based ambiguity set may be written as,

$$\min_{x \in X} \quad \max_{\mathbb{P}(\xi)} \quad \mathbb{E}_{\mathbb{P}(\xi)}\Big[f(x,\xi)\Big] \tag{3.15a}$$

$$\text{s.t.} \quad W_d(\mathbb{P},\mathbb{P}_0) \leq \epsilon_d \tag{3.15b}$$

where the loss function $f(x,\xi)$ is assumed to be an affine function of ξ. The tractable form of the W_d-DRO problem may be obtained by taking the dual of inner maximization problem in Model 3.15. First, the constraint 3.15b is replaced by the OT-GMM problem (Model 3.11) as follows,

$$\min_{x \in X} \quad \max \quad \mathbb{E}_{\mathbb{P}(\xi)}\Big[f(x,\xi)\Big] \tag{3.16a}$$

$$\text{s.t.} \quad \min_{\pi \geq 0} \quad \sum_{l=1}^{L}\sum_{l'=1}^{L} c_{l,l'}\pi_{l,l'} \leq \epsilon_d^2 \tag{3.16b}$$

$$\sum_{l'=1}^{L} \pi_{l,l'} = w_l^0, \quad \forall 1 \leq l \leq L \tag{3.16c}$$

After dropping the *min* operator in constraint 3.16b, the model is reformulated as,

$$\min_{x \in X} \quad \max_{\pi \geq 0} \quad \mathbb{E}_{\mathbb{P}(\xi)}\Big[f(x,\xi)\Big] \tag{3.17a}$$

$$\text{s.t.} \quad \sum_{l=1}^{L}\sum_{l'=1}^{L} c_{l,l'}\pi_{l,l'} \leq \epsilon_d^2 \tag{3.17b}$$

$$\sum_{l'=1}^{L} \pi_{l,l'} = w_l^0, \quad \forall 1 \leq l \leq L \tag{3.17c}$$

Since the candidate distributions $\mathbb{P}(\xi) \in \mathcal{P}$ are treated as GMMs, the expectation operator in the objective function (3.17a) may be expanded as

$$\mathbb{E}_{\mathbb{P}(\xi)}\Big[f(x,\xi)\Big] := \sum_{l'=1}^{L} w_{l'}\mathbb{E}_{\nu_{l'}^0}\Big[f(x,\xi)\Big] \tag{3.18a}$$

$$= \sum_{l=1}^{L} \sum_{l'=1}^{L} \pi_{l,l'} \mathbb{E}_{\nu_{l'}^0} \Big[f(x, \xi) \Big] \tag{3.18b}$$

$$= \sum_{l=1}^{L} \sum_{l'=1}^{L} \pi_{l,l'} f\Big(x, \mathbb{E}_{\nu_{l'}^0}[\xi] \Big) \tag{3.18c}$$

Note that the last step in the above is based on the assumption that the loss function is affine with respect to ξ. Substituting expression 3.18c into 3.17a, the model is reformulated as,

$$\begin{aligned}
\min_{x \in X} \quad \max_{\pi \geq 0} \quad & \sum_{l=1}^{L} \sum_{l'=1}^{L} \pi_{l,l'} f\big(x, \mathbb{E}_{\nu_{l'}^0}[\xi]\big) && (3.19a)\\
\text{s.t.} \quad & \sum_{l=1}^{L} \sum_{l'=1}^{L} c_{l,l'} \pi_{l,l'} \leq \epsilon_d^2 && (3.19b)\\
& \sum_{l'=1}^{L} \pi_{l,l'} = w_l^0, \quad \forall 1 \leq l \leq L && (3.19c)
\end{aligned}$$

The dual of the inner maximization problem in Model 3.19 may be found to give the overall minimization problem pertaining to the tractable W_d-DRO problem as follows,

$$\begin{aligned}
\min_{x, \eta \geq 0, y_l} \quad & \eta \epsilon_d^2 + \sum_{l=1}^{L} w_l^0 y_l && (3.20a)\\
\text{s.t.} \quad & y_l \geq f\big(x, \mathbb{E}_{\nu_{l'}^0}[\xi]\big) - \eta c_{l,l'}, \quad \forall 1 \leq l \leq L, 1 \leq l' \leq L && (3.20b)\\
& x \in X && (3.20c)
\end{aligned}$$

Here, η and y_l refer to the dual variables involved in the inner maximization problem, while x refers to the original decision variable for the problem. For an affine loss function, Model 3.20 is a linear programming (LP) problem. It may be noted, however, that the proposed W_d-DRO approach in Model 3.20 can also be applied to non-affine loss functions in ξ.

3.4 Numerical example

In this section, the results of the proposed W_d-DRO approach are compared and contrasted with those of the Wasserstein DRO problem through a numerical example. The DRO model under objective function uncertainty is taken as,

$$\min_{x_1,x_2} \quad \max_{\mathbb{P}(\xi)\in\mathcal{P}} \quad \mathbb{E}_{\mathbb{P}(\xi)}\big[x_1(1+\xi)+x_2\big] \tag{3.21a}$$

$$\text{s.t.} \quad 3.2x_1 + 0.2x_2 \geq 7.5 \tag{3.21b}$$

$$2x_1 + 3x_2 \geq 12 \tag{3.21c}$$

$$x_1 - 1.5 \leq x_2 \tag{3.21d}$$

$$12 - 4x_1 \geq -2x_2 \tag{3.21e}$$

$$-1.2x_1 + x_2 \leq 1 \tag{3.21f}$$

$$-0.3x_1 + 1.7x_2 \leq 6.2 \tag{3.21g}$$

$$x_1 + x_2 \leq 8 \tag{3.21h}$$

$$4.6x_1 + 5.2x_2 \geq 23.92 \tag{3.21i}$$

$$2.8x_1 + 0.76x_2 \geq 8.4 \tag{3.21j}$$

$$x_1, x_2 \geq 0 \tag{3.21k}$$

In this example, 500 instances of ξ are sampled (Figure 3.4) from a 3-component GMM, denoted as $\mathbb{P}^{\text{true}}$, with the following attributes

$$\mathbb{P}^{\text{true}} := \quad 0.24\mathcal{N}(-5.21, 2.39) + 0.4\mathcal{N}(1.23, 4.12) + 0.36\mathcal{N}(7.47, 3.36) \tag{3.22}$$

For the Wasserstein DRO problem, sampled uncertainty $\xi_k^0, \ \forall k = \{1, ..., 500\}$ is directly used in Model 3.6c to obtain the solution. However, for the W_d-DRO problem, the sampled uncertainty first needs to be fitted to a GMM. In this work, we fit the GMMs using MATLAB's *fitgmdist* function which utilizes the EM algorithm to obtain the GMM attributes. In this case, since 3 observable modes are present in the

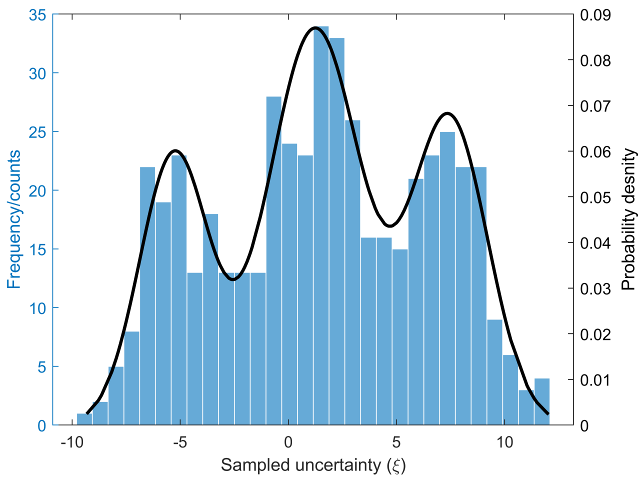


Figure 3.4: An illustration of sampled uncertainty from the source distribution (Equation 3.22) for the numerical example under distributional uncertainty in Model 3.21. The blue bars represent the histogram of the sample set, while the black line depicts the fitted GMM ($\mathbb{P}_0$).

histogram of the sampled uncertainty (Figure 3.4), a 3-component GMM is fitted, denoted as $\mathbb{P}_0$ to the samples as follows,

$$\mathbb{P}_0 := \; 0.24\mathcal{N}(-5.28, 2.56) + 0.46\mathcal{N}(1.23, 4.55) + 0.3\mathcal{N}(7.44, 3.16) \tag{3.23}$$

Then, the attributes of the Gaussian components in $\mathbb{P}_0$, namely the mean and variance of each components, are used in Model 3.20. Both the Wasserstein DRO problem, as well as the W_d-DRO problem were solved for increasingly large radii of the ambiguity set (ϵ_W for Wasserstein DRO, and ϵ_d for W_d-DRO) to obtain the optimal solution, as well as the DRO optimal objective. The results of the Wasserstein, and W_d-DRO problems are shown in Table 3.1.

From Figure 3.5, we observe that the Wasserstein DRO optimal objective increases sharply with an increase in the ambiguity set radius ϵ_W. In contrast, the W_d-DRO optimal objective shows a much less sharply increasing trend with an increase in the ambiguity set radius ϵ_d. The reason for this difference in the trends can be explained by highlighting the main difference between the approaches, namely the ambiguity set construction step. As mentioned earlier, the Wasserstein DRO approach hedges

Table 3.1: Evolution of optimal model objective values through Wasserstein, and W_d-DRO methods as a function of ambiguity set radius (ϵ)

Ambiguity	DRO optimal objective value	
set radius	Wasserstein	W_d
(ϵ_W or ϵ_d)	DRO	DRO
0.01	8.44	8.40
1	11.85	8.77
10	42.86	21.27
100	352.95	21.27

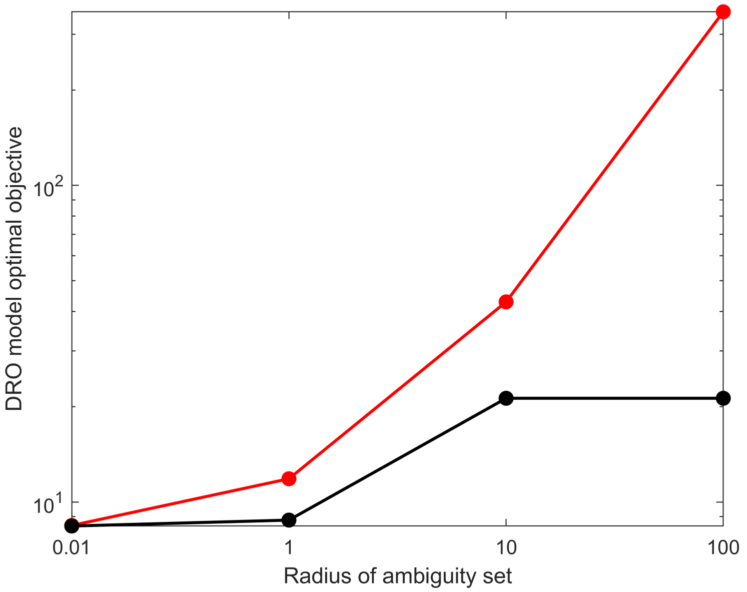


Figure 3.5: Evolution of the DRO model optimal objective value with radius of the ambiguity set for Wasserstein DRO (red) and the proposed W_d-DRO (black) for the numerical study

against all such candidate distributions $\mathbb{P}$ in its ambiguity set that present within the threshold radius ϵ_W of the nominal distribution $\mathbb{P}_0$. However, this method does not impose any further restrictions on the type of candidate distributions to be included in its ambiguity set. Therefore, for Wasserstein DRO, at large radii, there is a distinct possibility that the ambiguity set might be too rich, thus giving a very conservative optimal objective. In contrast, the W_d-DRO approach restricts its ambiguity set to contain only GMMs, specifically those GMMs based on the same components as those of the nominal distribution $\mathbb{P}_0$. Therefore, for W_d-DRO, even at larger radii, the ambiguity set is restricted to a smaller number of distributions than the Wasserstein DRO. Furthermore, this restriction on the distributions included in the ambiguity set is not arbitrary; rather, it is a restriction that is informed through the fitting of the nominal GMM $\mathbb{P}_0$ (Equation 3.23) to the sampled uncertainty, thus ensuring that information about the subpopulations or modes present in the samples is incorporated into the optimization step.

For all the studies chosen in this work, the Wasserstein and Wd DRO methods have a comparable computational time to solve in GAMS on a desktop computer with Intel(R) Core(TM) i5-4570 CPU @ 3.20GHz, 3201 Mhz, 4 Core(s). Note that, for the numerical example, the Wasserstein (QCP) DRO problem and the Wd (LP) problem were both solved using XPRESS with around 0.5 second; for the case studies, the Wasserstein and Wd DRO problems (both QCP owing to the second term in the loss function) were also solved using XPRESS.

3.4.1 Comparing the worst case distribution

As mentioned in Section 3.2.1, the DRO approach to a problem under uncertainty aims to find the optimal solution by hedging against the expectation taken over the worst case distribution, further denoted by $\mathbb{P}_{\text{wc}}$ in this chapter, in the ambiguity set. To this end, in this section, we compute and illustrate the worst case distribution for Wasserstein DRO, as well as W_d-DRO, for different radii of the ambiguity set.

The worst case distributions for different ambiguity set radii, denoted as $\mathbb{P}_{\text{wc}}|_{\epsilon_W}$ and $P_{\text{wc}}|_{\epsilon_d}$ respectively for Wasserstein DRO and W_d-DRO, are found as follows: for a specific ambiguity set radius, the DRO problem is solved to obtain the optimal solution x^*; this x^* is further used to solve the inner maximization problem in the

DRO model (Model 3.1) to obtain the probability masses of the elements supporting the candidate distribution, which in this case, is the worst case distribution.

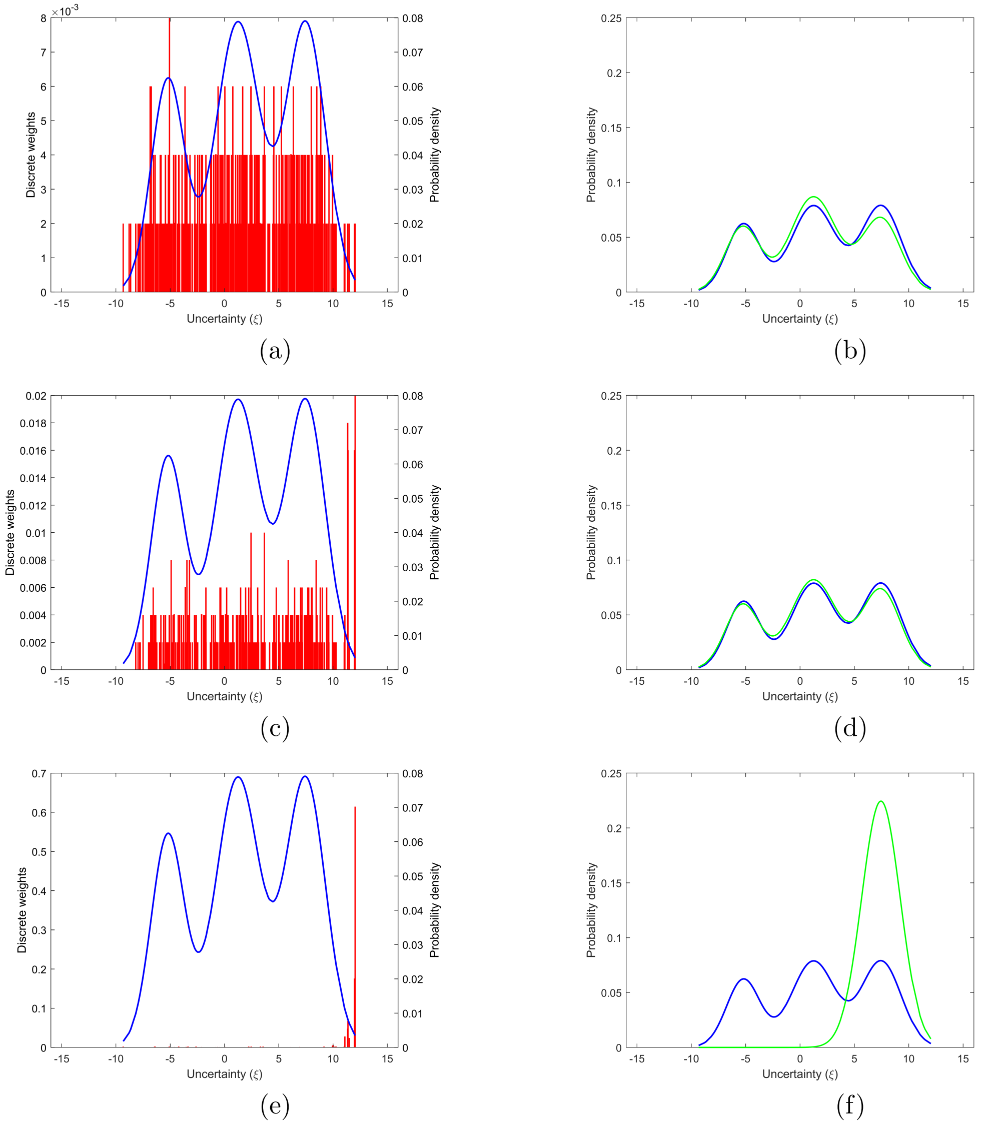
Table 3.2 presents the GMMs that give the worst case expectation of the loss function for W_d-DRO for various radii ϵ_d ranging from 0.01 to 100, centered around the same nominal distribution ($\mathbb{P}_0$), which is the GMM fitted to the sampled uncertainty (Equation 3.23). At smaller radii, such as 0.01 ~ 1, the worst case probability distribution at the periphery of the ambiguity set is found to be a GMM ($\mathbb{P}_{\mathrm{wc}}|_{\epsilon_d=0.01}$,$\mathbb{P}_{\mathrm{wc}}|_{\epsilon_d=1}$) whose component weights are close to those of the nominal GMM. When the radius is increased to 3.5, 5.3 and 6.5, the worst case distribution is a GMM based on the same components as $\mathbb{P}_0$, but with markedly different weights. When the radius is increased to larger values such as 10 ~ 100, the worst case GMM tends to a single Gaussian component among the 3 components in $\mathbb{P}_0$. The worst case distributions for ambiguity set radii $\epsilon_d \coloneqq \{0.01, 1, 3.5, 5.3, 6.5, 10, 100\}$ are listed in Table 3.2.

Table 3.2: Worst case distribution under different ambiguity set radius. Numbers shown in the table are the weighting coefficients of the GMM distribution.

ϵ_d	$\mathcal{N}(-5.283, 2.564)$	$\mathcal{N}(1.232, 4.553)$	$\mathcal{N}(7.442, 3.159)$
0.01	0.238	0.464	0.298
1	0.238	0.438	0.324
3.5	0.238	0.148	0.615
5.3	0	0.238	0.762
6.5	0	0.119	0.881
10	0	0	1
100	0	0	1

Figures 3.6a - 3.6f illustrate the worst case distributions for ambiguity set radii 0.01, 1, and 10. It must be noted that in the case of Wasserstein DRO, the problem hedges against a *discrete* worst case distribution that is supported on at most $N + 1$ points, where N refers to the size of the sample set (Yue *et al.* 2022). Therefore, the

DRO solution obtained may suffer from the issue of over-conservativeness since it is may hedge against unnecessary distributions. In contrast, for W_d-DRO, the worst case distribution is always a continuous distribution; at low values of ϵ_d, $\mathbb{P}_{\text{wc}}$ tends to the nominal GMM ($\mathbb{P}_0$), whereas for higher values of ϵ_d, the worst case distribution tends to a singular component of $\mathbb{P}_0$.



(a)

(b)

(c)

(d)

(e)

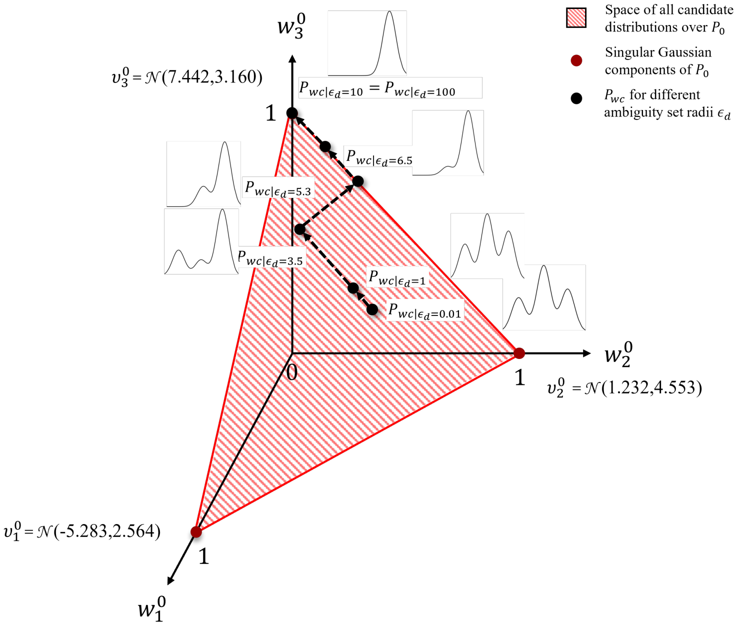
(f)

Figure 3.6: Illustrations of the Wasserstein worst-case distribution vs the W_d worst-case distribution, respectively, for radii of 0.01 [(a) and (b)], 1 [(c) and (d)], and 10 [(e) and (f)], respectively. The red bars represent the discrete weights of the support set of the worst case expectation distribution in Wasserstein DRO, the blue curve depicts the true/source distribution in Equation 3.22, and the green curve represents the worst case expectation distribution in W_d-DRO.

3.4.2 Upper bound on ϵ_d for W_d-DRO

As mentioned in Section 3.3, in the W_d-DRO approach, all candidate distributions in the ambiguity set are assumed to be based on the same components as the nominal GMM ($\mathbb{P}_0$). Figure 3.7 illustrates the simplex that may be constructed given $\mathbb{P}_0$, whose edges contain all countably infinite possible combinations of the Gaussian components in $\mathbb{P}_0$. The vertices of this simplex, namely ν_1^0, ν_2^0 and ν_3^0, correspond to the three fitted Gaussian components in $\mathbb{P}_0$. From the aforementioned assumption, and owing to the existence of the optimal solution of a linear programming problem at the vertex of its feasible region, an upper bound on the ambiguity set radius ϵ_d may be obtained as

$$\epsilon_d \leq \max_{1 \leq l \leq L} \; W_d(\mathbb{P}_0, \nu_l^0) \tag{3.24}$$



Space of all candidate distributions over P_0
Singular Gaussian components of P_0
P_{wc} for different ambiguity set radii ϵ_d
w_3^0
$\upsilon_3^0 = \mathcal{N}(7.442,3.160)$
$P_{wc|\epsilon_d=10} = P_{wc|\epsilon_d=100}$
1
$P_{wc|\epsilon_d=6.5}$
$P_{wc|\epsilon_d=5.3}$
$P_{wc|\epsilon_d=3.5}$
$P_{wc|\epsilon_d=1}$
$P_{wc|\epsilon_d=0.01}$
w_2^0
0
1
$\upsilon_2^0 = \mathcal{N}(1.232,4.553)$
$\upsilon_1^0 = \mathcal{N}(-5.283,2.564)$
1
w_1^0

Figure 3.7: An illustration of the worst-case distributions ($\mathbb{P}_{\mathrm{wc}}$) for different ambiguity set radii (ϵ_d)

For the numerical study in Section 3.4, the upper bound on ϵ_d, denoted as $\mathcal{U}[\epsilon_d]$, using Equation 3.24 is as follows

$$\mathcal{U}[\epsilon_d] := \max\left\{W_d(\mathbb{P}_0,\nu_1^0), W_d(\mathbb{P}_0,\nu_2^0), W_d(\mathbb{P}_0,\nu_3^0)\right\} = \max\left\{8.253, 4.657, 7.513\right\} = 8.253 \tag{3.25}$$

Here, $W_d(\mathbb{P}_0,\nu_l^0),\ \forall 1 \leq l \leq L$ is computed using Model 3.11, where $W_d := \sqrt{\delta}$. It must be noted that choosing $\mathcal{U}[\epsilon_d] = 8.253$ does not imply that the worst case distribution for the W_d-DRO problem is $\nu_1^0 := \mathcal{N}(-5.283, 2.564)$. Rather, when the W_d-DRO problem is solved for ambiguity set radius $\epsilon_d = \mathcal{U}[\epsilon_d] = 8.253$, and the inner maximization problem in Model 3.16 solved using the resulting DRO optimal solution $(x_1^*, x_2^*)|_{\epsilon_d=8.253} = (2.147, 3.141)$, the worst case distribution $\mathbb{P}_{\text{wc}}$ is obtained as

$$\mathbb{P}_{\text{wc}}|_{\epsilon_d=8.253} := \mathcal{N}(7.442, 3.159) \tag{3.26}$$

That is, $\mathbb{P}_{\text{wc}}|_{\epsilon_d=8.253}$ is the worst case distribution corresponding to the largest possible ambiguity set radius (computed as $\mathcal{U}[\epsilon_d]$), for the fitted $\mathbb{P}_0$. Upon solving the optimal transport problem between $\mathbb{P}_0$ and $\mathbb{P}_{wc}|_{\epsilon_d=8.253}$, it was found that $W_d(\mathbb{P}_0, \mathbb{P}_{\text{wc}}|_{\epsilon_d=8.253}) = 7.513$, thus illustrating that the computed value of 8.253 is indeed an upper bounding value on 7.513. It is further observed, and illustrated in Figure 3.7, that increasing ϵ_d beyond 7.513 to larger values such as 10 ~ 100 does not change the DRO solution for a given $\mathbb{P}_0$.

3.5 Case study - a portfolio optimization example

Having illustrated some features of the proposed W_d-DRO approach, and contrasted its performance on a numerical study in Section 3.4, we further applied the proposed method to a portfolio optimization case study.

The portfolio optimization case study in this section is adapted from the work of Esfahani and Kuhn (2018). In their study, they designed a mean-risk portfolio optimization problem that minimizes the weighted sum of the mean and the conditional

Value-at-Risk (CVaR) of the portfolio loss. In our study, we chose to replace the CVaR measure with the variance of the optimal portfolio to get the following DRO problem,

$$\min_{x\in\mathbb{R}^m} \quad \max_{\mathbb{P}(\xi)\in\mathcal{P}} \quad \mathbb{E}_{\mathbb{P}(\xi)}[-\xi^\intercal x] + \lambda x^\intercal \Sigma x \tag{3.27a}$$

$$\text{s.t.} \quad \mathbf{1}^\mathrm{T} x = 1 \tag{3.27b}$$

Here, the portfolio loss is denoted by $-\xi^\intercal x$ where x_i denotes the fraction of each asset $i = 1, ..., m$ in the portfolio. The uncertainty in the problem arises from the returns ξ_i on each asset. The assets are sorted such that lower-indexed assets offer smaller mean returns, while higher indexed assets offer larger mean returns but with relatively higher variance. For this study, we set $m = 10$. Since the portfolio variance term in the objective function is independent of uncertain returns ξ, it is brought out of the inner maximization problem such that the loss function addressed in the DRO problem is linear with respect to ξ.

Using the formulations described in Models 3.6 and 3.20, respectively, we obtained the Wasserstein DRO model specific to the portfolio optimization case study as,

$$\min_{x,\eta\geq 0,z_k} \quad \eta\epsilon_W + \frac{1}{N}\sum_{k=1}^{N} z_k + \lambda \sum_{i=1}^{m} x_i \sum_{i'=1}^{m} \sigma_{i',i} x_{i'} \tag{3.28a}$$

$$\text{s.t.} \quad \sum_{i=1}^{m} -x_i \xi_{k,i}^0 \leq z_k, \quad 1 \leq k \leq N \tag{3.28b}$$

$$\sum_{i=1}^{m} x_i^2 \leq \eta^2 \tag{3.28c}$$

$$\sum_{i=1}^{m} x_i = 1 \tag{3.28d}$$

and the corresponding W_d-DRO model is

$$\min_{x,\eta\geq 0,y_l} \quad \eta\epsilon_d^2 + \sum_{l=1}^{L} w_l^0 y_l + \lambda \sum_{i=1}^{m} x_i \sum_{i'=1}^{m} \sigma_{i',i} x_{i'} \tag{3.29a}$$

$$\text{s.t.} \quad \sum_{i=1}^{m} -x_i \mathbb{E}_{\nu_{l'}^0}[\xi_i] - \eta c_{l,l'} \leq y_l, \quad 1 \leq l, l' \leq L \tag{3.29b}$$

$$\sum_{i=1}^{m} x_i = 1 \tag{3.29c}$$

3.5.1 Gaussian mixture distribution-based uncertainty

To test the performance of the proposed method, we assume that the uncertainty is sourced from a 3-component Gaussian mixture model as follows, wherein the Gaussian components are described using their corresponding means and variances,,

$$\xi_i \sim 0.6\mathcal{N}\Big(i \times 0.75, i \times 1.8\Big) + 0.25\mathcal{N}\Big(i \times 2.5, i \times 1.8\Big) + 0.15\mathcal{N}\Big(i \times 4.25, i \times 1.8\Big) \tag{3.30}$$

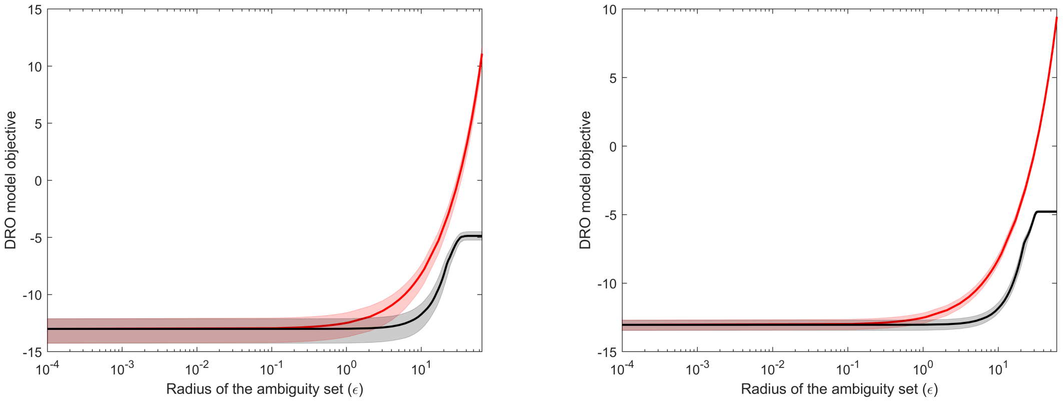
We used the data directly to solve the Wasserstein DRO model. Prior to solving W_d-DRO, we first fit the sampled uncertainty to a GMM whose component number we specify. In this study, we fit the samples to a 3-component nominal GMM distribution ($\mathbb{P}_0$). For the purpose of studying the effect of sample size on the DRO performance, we solved the models for two sample set sizes $N = \{50, 500\}$. Furthermore, to assess the DRO performance from the lens of in-sample as well as out-of-sample stability, we performed 200 independent simulations; that is, we generated 200 independent datasets of samples and obtained their respective resulting optimal portfolios (x^*) to test for performance.

In this section, we discuss the performance of the Wasserstein and W_d-DRO problems in the context of in-sample and out-of-sample stability of their respective optimal solutions trained on the same dataset, as well as the reliability of their solutions. In-sample stability refers to how stable the solution from a proposed model is for different datasets generated from the same source. In contrast, out-of-sample stability refers to how stable the expected returns are when the solutions from the proposed models are used in conjunction with a large test sample dataset. Therefore, in-sample stability provides a measure of how sensitive a model is to a difference in the sample data, while out-of-sample stability is a better measure of the actual model performance itself (Kaut and Stein 2003). In this study, we assess the solutions by computing their reliability (Esfahani and Kuhn 2018) across an increasing range of ambiguity set radii, and further compare the Wasserstein and W_d-DRO solutions at a defined threshold of 95%.

Figures 3.8a and 3.8b illustrate the in-sample stability results, namely the DRO model objective as a function of ambiguity set radius, for the portfolio optimization case study for uncertain asset returns sampled from a 3-component GMM, as well as the reliability of the solutions. From the figures, as illustrated in Section 3.4 also, it is observed that as the radius of the ambiguity set radius increases, the Wasserstein DRO (red) model objective increases steeply, while for W_d-DRO (black), the increase is less steep, and saturates beyond a certain radius. In the context of variability of the model objective, it is observed that both Wasserstein and W_d-DRO display comparable in-sample model objective variability for smaller ambiguity set radii, while for larger radii, the variability of the W_d-DRO method is slightly more than that of Wasserstein DRO. It is also observed that an increase in number of samples (in this case, from 50 to 500) results in an increase in in-sample stability for both DRO methods.

Figures 3.9a and 3.9b showcase the out-of-sample stability results of the Wasserstein (red) and W_d-DRO (black) methods for 50, and 500 samples each. In the figures, the solid lines depict the average values out of 200 trials, while the shaded region represents the area between the 20% and 80% percentiles, respectively. The dotted lines depict the evolution of reliability of the Wasserstein (red) and W_d-DRO (black) solutions for increasing ambiguity set radii. The green solid line depicts the 95% reliability threshold.

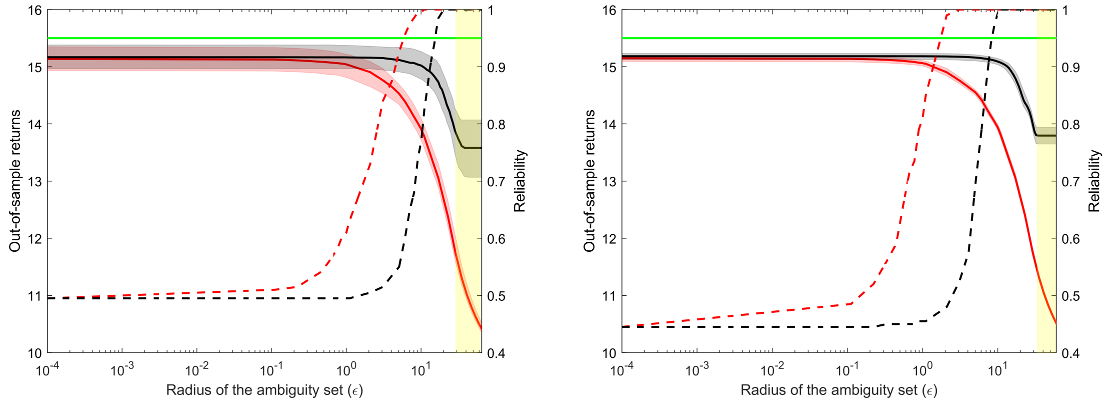
From the results illustrated in Figure 3.9, the following observations and inferences



(a)

(b)

Figure 3.8: Evolution of in-sample model objective through Wasserstein DRO (red) and W_d-DRO (black) with respect of ambiguity set radius for (a) 50 samples, and (b) 500 samples of uncertainty sourced from a 3-component GMM; the solid lines represent the means and the shaded region represents the 20% and 80% percentile values over 200 independent trials.



(a)

(b)

Figure 3.9: Evolution of expected out-of-sample returns through Wasserstein DRO (red) and W_d-DRO (black) with respect of ambiguity set radius for (a) 50 samples, and (b) 500 samples of uncertainty sourced from a 3-component GMM; the solid lines represent the means and the shaded region represents the 20% and 80% percentile values over 200 independent trials. The yellow highlighted region depicts the range of the upper bounding radii calculated for the W_d-DRO method, as put forth in Section 3.4.2, across the 200 independent trials. The green solid line depicts the 95% reliability threshold, while the dotted lines represent the evolution of reliability of the Wasserstein (red) and W_d-DRO (black) methods.

can be made specifically regarding how the models perform in a test scenario.

- Firstly, we discuss the average performance of the expected out-of-sample returns using the solutions obtained via the Wasserstein and W_d-DRO methods. For smaller ambiguity set sizes, both Wasserstein DRO as well as W_d-DRO offer a comparable performance. As the radius of the ambiguity set increases, Wasserstein DRO's average out-of-sample performance appears to deteriorate steeply; in contrast, W_d-DRO's performance only shows a slight dip in returns and saturates beyond a certain radius.

- Secondly, we discuss the effect of sample set size on the performance of both Wasserstein and W_d-DRO. With a larger number of samples, it is found that the variability of expected out-of-sample returns of Wasserstein as well as W_d-DRO, depicted by the shaded red and black regions respectively, decreases significantly. In the case of W_d-DRO, this may be attributed to the ability to find a better nominal GMM that describes the sampled uncertainty when a larger number of samples is available.

- We assess the quality of the obtained Wasserstein and W_d-DRO solutions through the lens of reliability. From Figures 3.9a and 3.9b, we see that for 95% reliability, the W_d-DRO method offers larger average returns than the Wasserstein method.

- Finally, we draw the reader's attention to the highlighted region (yellow) in Figures 3.9a and 3.9b. This region depicts the various upper bounding radii obtained via the approach described in Equation 3.24, for the nominal GMMs fitted to the samples corresponding to the 200 independent trials. As previously discussed, increasing the ambiguity set radius beyond a certain value does not result in a change in the worst case expectation distribution for W_d-DRO. This is because the W_d-DRO problem formulation is designed such that the candidate distributions are described on the same Gaussian components as the nominal distribution, and the worst case expectation distribution may be reasonably assumed to tend towards a single Gaussian component (refer to Figure 3.7. Simply put, the highlighted range of ambiguity set radii in Figure 3.9 correspond to a good enough ambiguity set size for the datasets under consideration to give 100% reliable solutions. This result illustrates an advantage over the conventional Wasserstein DRO, which requires the user to try different radii and check whether desired reliability is achieved.

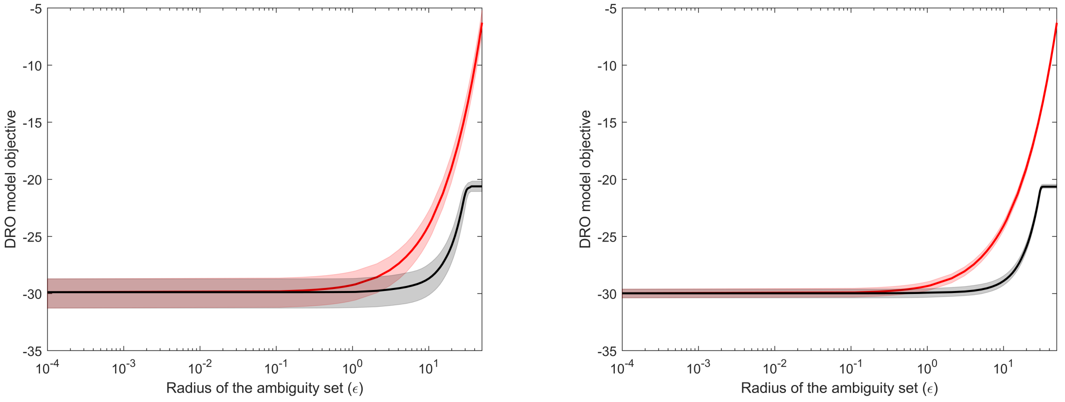
3.5.2 Lognormal mixture distribution-based uncertainty

In this section, we study the case wherein the true uncertainty distribution is a lognormal mixture distribution, instead of GMM distribution, to test the performance of the proposed method when the true distribution is not a perfect GMM. We define a mixture (x) of two Gaussian distributions, and define ξ as the exponential of the mixture, leading to a lognormal mixture distribution as follows,

$$x_i \sim 0.6\mathcal{N}\Big(i \times 2.5, i \times 1.8\Big) + 0.4\mathcal{N}\Big(i \times 5, i \times 1.8\Big) \tag{3.31}$$

$$\xi_i \sim \exp(x_i) \tag{3.32}$$

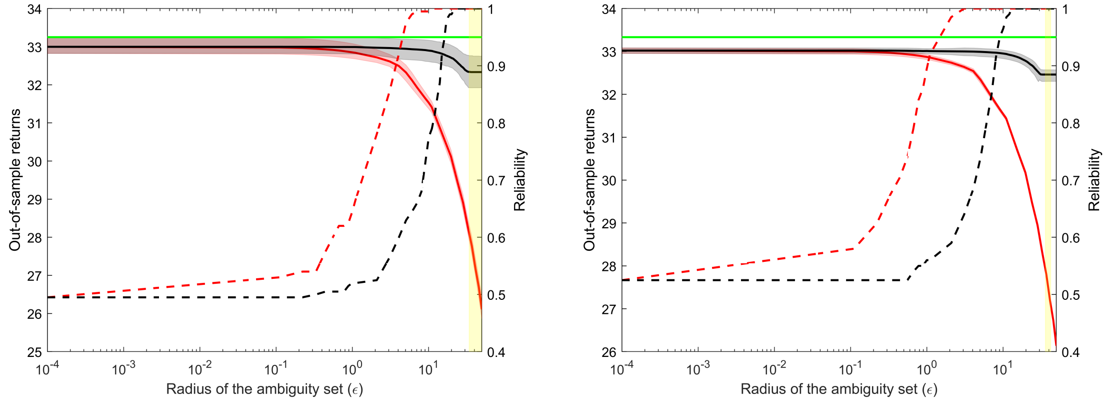
Figures 3.10a and 3.10b depict the in-sample stability results, while Figures 3.11a and 3.11b depict the out-of-sample performance of the Wasserstein and W_d-DRO models, respectively. From Figures 3.10a and 3.10b, we find that in-sample stability improves for both methods with an increase in sample size, as well as with an increase in ambiguity set radius. From figures 3.11a and 3.11b, we see that the average out-of-sample returns from the W_d-DRO method are higher than those from the Wasserstein method for samples sourced from a lognormal mixture distribution. These results show that the improved performance of the proposed W_d-DRO method is not just limited to sampled uncertainty sourced from a Gaussian mixture distribution. Rather, the ability of a Gaussian Mixture Model to function as a universal density estimator (Aragam *et al.* 2018) enables the fitting of the multimodal distribution to a GMM, thus constructing an ambiguity set with a tighter selection of candidate distributions leading to lesser conservative solutions with higher out-of-sample expected returns.



(a)

(b)

Figure 3.10: Evolution of in-sample model objective through Wasserstein DRO (red) and W_d-DRO (black) with respect of ambiguity set radius for (a) 50 samples, and (b) 500 samples of uncertainty sourced from a lognormal mixture distribution; the solid lines represent the means and the shaded region represents the 20% and 80% percentile values over 200 independent trials.



(a)

(b)

Figure 3.11: Evolution of expected out-of-sample returns through Wasserstein DRO (red) and W_d-DRO (black) with respect of ambiguity set radius for (a) 50 samples, and (b) 500 samples of uncertainty sourced from a lognormal mixture distribution; the solid lines represent the means and the shaded region represents the 20% and 80% percentile values over 200 independent trials. The yellow highlighted region depicts the range of the upper bounding radii calculated for the W_d-DRO method, as put forth in Section 3.4.2, across the 200 independent trials. The green solid line depicts the 95% reliability threshold, while the dotted lines represent the evolution of reliability of the Wasserstein (red) and W_d-DRO (black) methods.

3.6 Summary

In this work, we presented a novel distributionally robust optimization method, that uses a recently-developed variant of optimal transport distance between Gaussian Mixture Models. Since the cost function associated in this problem is computed using the closed-form expression for the 2-Wasserstein metric between Gaussian distributions, it is possible to incorporate statistical information such as the mean and variance, namely the first and second moments, into the optimal transport metric-based DRO problem. We presented the tractable formulation of the DRO problem under the OT-GMM-based ambiguity set, and illustrated its use on a numerical example. Through this example, we discussed the evolution of the DRO optimal objective through our method, in contrast to the established Wasserstein DRO approach. We also presented results on the worst-case expectation distribution in the proposed DRO method, and presented a simple method to compute an upper bound on the radius of the ambiguity set which is an important hyperparameter in DRO that affects the conservativeness of the solution. We further conducted a portfolio optimization case study using the proposed DRO method, and examined the out-of-sample performance of the method for uncertainty sampled from two types of source distributions. Through our work, we observed that the proposed OT-GMM-based DRO method offers significant improvement in the out-of-sample performance, as compared to Wasserstein DRO. We also note that the efficacy of the proposed method is dependent on the quality of the GMM fitted to the available data. In this sense, to have sufficient amount of data for GMM fitting is a requirement for the proposed method.

The DRO method using a GMM-based ambiguity set provides a method to assess the worst-case performance of an optimal solution that was obtained by quantifying a level of confidence about the uncertain parameters' probability distribution into the optimization problem by means of an ambiguity set. In many practical applications, it is preferable to build upon this “worst-case” expectation problem, and attempt to minimize the worst-case performance. One such method that has been developed in literature is the distributionally robust chance-constrained programming method (DRCCP) wherein the worst-case expected value of a probabilistic constraint in an optimization model is constrained using a threshold. In Chapter 4, we use the worst-case expectation problem under a GMM-based ambiguity set developed in this chapter to address this DRCCP problem.

Chapter 4

Distributionally Robust Chance-Constrained Optimization with a Gaussian Mixture Optimal Transport-based Ambiguity Set

Abstract: Conventional chance-constrained programming methods suffer from the inexactness of the estimated probability distribution of the underlying uncertainty from data. To this end, a distributionally robust approach to the problem allows for a level of ambiguity considered around a reference distribution. In this work, we propose a novel formulation for the distributionally robust chance-constrained programming problem using an ambiguity set constructed from a variant of optimal transport distance that was developed for Gaussian Mixture Models. We show that for multimodal process uncertainty, our proposed method provides an effective way to incorporate statistical moment information into the ambiguity set construction step, thus leading to improved optimal solutions. We illustrate the performance of our method on a numerical example as well as a chemical process case study. We show that our proposed methodology leverages the multimodal characteristics from the uncertainty data to give superior performance over the traditional Wasserstein distance-based method.

4.1 Introduction

Real industrial and allied processes are often fraught with a number of uncertain factors ranging from demand and supply, pricing, process measurements and model parameter estimation. In such cases, any optimization model or application developed must take into account these uncertainties in order to give not only a mathematically optimal solution, but also a practically feasible one (Sahinidis 2004; Ning and You 2019). Optimization under uncertainty has been a topic of great academic and industrial interest since the late 1900s wherein mathematical programming techniques have been used in areas such as process design, planning and scheduling operations, as well as process control. The main methods of optimization under uncertainty are distinguished primarily by their treatment of the probability distribution of the underlying uncertainty in the process. Stochastic programming methods consider that the probability distribution of the uncertain parameters is known, or that it can be well-estimated from the available data, and incorporates this distributional information into its formulations (Dantzig 1955; Birge and Louveaux 2011). Robust optimization methods, on the other hand, disregard any information on the probability distribution of the uncertainty and instead optimize for the worst-case realization of the problem subject to an uncertainty set (Ben-Tal and Nemirovski 1998; Ben-Tal *et al.* 2009). A considerable amount of research and applications have been undertaken using variants of stochastic and robust optimization techniques.

From a theoretical standpoint, operations research focusing on decision theory makes a distinction between the ideas of “risk” and “ambiguity” associated with an optimization problem (Keynes 2013). Simply put, risk points to underlying uncertainty accounted for in the problem by means of a known probability distribution, while ambiguity refers to the uncertainty in the knowledge of this distribution itself (Wiesemann *et al.* 2014). Stochastic programming and robust optimization approaches to decision-making under uncertainty address the incorporation of the underlying risk in the process into the optimization problem; however, neither class of methods is well-equipped to deal with ambiguity. Recent research has focused on an intermediate approach to stochastic programming and robust optimization methods by considering "distributional ambiguity". Distributionally robust optimization (DRO), also known as ambiguous stochastic optimization, considers limited distributional information. It safeguards against worst-case outcomes within an "ambiguity set" of candidate distributions centered around a nominal distribution. This nominal distribution may be obtained from domain knowledge, as well as statistical analysis of the process history data. In the context of real-world performance of optimal solutions obtained through various methods, it has been empirically proven that accounting for distributional robustness is favorable since hedging against a single probability distribution of uncertainty often leads to poor test (or out-of-sample) performance (Esfahani and Kuhn 2018).

Distributionally robust optimization (DRO) was first well-addressed in optimization literature by Scarf *et al.* (1957). In this work, the authors addressed worst-case profit maximization of an inventory under distributional ambiguity of the product demand with a known mean and variance. From a modeling perspective, DRO may be formulated as a semi-infinite programming problem with respect to the space of probability distributions in the ambiguity set. The methods dealing with the computational intractability arising from this semi-infinite aspect may be broadly classified into cutting plane approaches, and the dual method. In the former, the semi-infinite quantifier of the candidate distributions is approximated by a finite atomic subset of the space. In each iteration of the method, a new probability distribution is added to this finite approximation and the problem is solved until optimality criteria are met (Mehrotra and Papp 2014; Bansal *et al.* 2018). In contrast, the dual methods leverage linear, Lagrangian and conic duality principles in order to convert the original primal maximization over the continuous candidate space of probability distributions to a dual minimization problem (Bertsimas *et al.* 2010; Ben-Tal *et al.* 2013). These methods may be employed when strong duality holds, and find notable use when the ambiguity set formulation may itself be defined using the decision variables in the problem (termed "decision-dependent ambiguity sets") (Noyan *et al.* 2018).

Ambiguity sets put forth in literature may be classified into four main groups, namely, discrepancy-based, moment-based, shape-preserving, and kernel-based (Rahimian and Mehrotra 2019). As previously mentioned, an ambiguity set contains a number of probability distributions that share common features; these four groups differ in these features. Specifically, moment-based ambiguity sets contain those probability distributions that all satisfy certain statistical moment properties, while shape-preserving sets exhibit similar structural qualities such as symmetry, and skewness. Kernel-based ambiguity sets contain common distributions formed through a kernel function whose characteristics closely match those of a reference kernel function. It is important to note that this grouping of ambiguity sets is not disjoint, and that certain methods of ambiguity set construction may be classified into more than one of these groups. The "shape" and "size" of the ambiguity set are key descriptors of an ambiguity set. The shape of the ambiguity set plays into the tractability of the associated DRO model; for practical ease of implementation, researchers have focused on constructing ambiguity sets that give rise to "solvable" formulations, such as a linear programming (LP) form, a second-order conic (SOCP) form, or a semi-definite (SDP) form (Rahimian and Mehrotra 2019). The size of the ambiguity set, on the other hand, is associated with the level of ambiguity associated with the underlying probability distribution of uncertainty; in practice, this is a hyperparameter tuned to the process needs.

Discrepancy-based ambiguity sets are constructed around a reference or "nominal" distribution which is usually an empirical distribution estimated from the available process history data on uncertainty. As the name suggests, discrepancy-based ambiguity sets account for all those probability distributions whose similarity to the reference distribution may be quantified by a chosen "discrepancy measure" of a user-defined magnitude (which controls the size, ergo also the level of distributional robustness accounted for in the problem), which may or may not be a metric. A number of such measures have been studied in the context of DRO, such as optimal transport discrepancy measures/distances, ϕ–divergences, and L_p norms. It may be noted that certain discrepancy-based ambiguity sets are classified as metric-based sets (e.g., optimal transport distance), while others are not (e.g., Kullback-Leibler Divergence).

Optimal transport (OT) problem admits a proper metric in the space of probability distributions; furthermore, it does not restrict the distributions in the ambiguity set to share the same support as that of the nominal distribution (Villani *et al.* 2009). In recent years, a significant number of works have focused on the use of this metric for ambiguity set construction; one of the first works using OT/Wasserstein distance (as the Kantorovich distance) in this context was published by Pflug and Wozabal (2007) for a portfolio optimization problem under distributional ambiguity of stock returns associated with their assets. Some other notable works in this area utilizing the Wasserstein distance are summarized here. Blanchet and Murthy (2019) proposed a strong dual one-dimensional formulation of the worst-case expected value problem over the ambiguity set, while considering the nominal distribution supported on general Polish spaces. Esfahani and Kuhn (2018) model the nominal distribution as an empirical uniformly-weighted discrete measure, for specific structures of uncertainty-riddled functions in the optimization problem, and for norm-based optimal transport cost. An important contribution of this work is the reformulation of the DRO problem as a finite-dimensional convex problem; furthermore, the authors also detail a method to construct the worst-case distribution and present finite-sample as well as asymptotic consistency guarantees using a portfolio optimization case study. Gao and Kleywegt (2023) also utilized the $p-$Wasserstein metric to construct an ambiguity set for DRO, wherein the cost function admits a metric on a Polish space, and further provide strong duality results using Lagrangian duality and obtain the worst-case distribution (or its approximation) using first-order optimality conditions from their dual reformulation of the worst-case expectation problem. In addition to these works, a number of other works have studied various reformulations of the DRO problem under specific assumptions, such as a conic reformulation for distributionally robust two-stage stochastic programming problems (Hanasusanto and Kuhn 2018), and a semi-infinite programming problem illustrated on a logistic regression case study (Luo and Mehrotra 2019).

In distributionally robust chance constrained optimization problem, the constraints are affected by uncertainty without exact distribution information. The worst-case probability of chance-constraint satisfaction is constrained to a minimum threshold. A few works in literature have utilized various discrepancy measures to obtain tractable formulations. Jiang and Guan (2016) derived an exact reformulation of a data-driven stochastic programming problem with distributionally robust chance constraints with a $\phi-$divergence-based ambiguity set formulation, which they showed is equivalent to the classical form of a chance constraint with a shifted risk level. Xie and Ahmed (2018) studied distributionally robust chance-constrained programming problems (DRCCP) with convex nonlinear uncertain constraints, and a moment-based ambiguity set, and provided tractable convex reformulations for the problem under specific assumptions, as well as a tractable mixed-integer convex reformulation for such DRCCP involving binary variables. Ji and Lejeune (2021) proposed reformulation frameworks for DRCCP with individual as well as joint chance constraints, ranging from mixed integer linear programming to exact mixed integer second order cone programming frameworks depending on the type of uncertainty considered, and illustrated their use cases on a knapsack problem. Chen *et al.* (2022) studied data-driven DRCCPs with a general p-Wasserstein distance-based ambiguity set, and provided exact tractable reformulations for both individual as well as joint chance constrained problems with right-hand side uncertainty as mixed-integer conic problems; furthermore, they showed that for special cases of $p = 1$ or $p = \infty$, the reformulation tends to a mixed-integer linear problem.

Formulating a tractable approach and determining the best decisions hinge greatly on how we define the ambiguity set for Distributionally Robust Optimization (DRO). Designing this set is critical, as it should capture a reasonable level of uncertainty about the probability distribution while excluding irrelevant or extreme distributions that could skew decisions. This chapter delves into creating discrepancy-based ambiguity sets, encompassing distributions close to a nominal one according to a distance metric. Our focus lies on addressing uncertainty showcasing multi-modal distributional characteristics. Rather than employing a broad ambiguity set, we propose utilizing the optimal transport distance between Gaussian mixture models, tailored to capture relevant distributions. This approach aims to hedge against pertinent candidates. Building on this, we introduce a novel Distributionally Robust Chance-Constrained Programming (DRCCP) method utilizing a variant of optimal transport for Gaussian mixture models. Unlike existing techniques, our method extends to a broader range of uncertain constraint functions, not limited by linearity or convexity constraints.

The rest of the chapter is organized as follows. In Section 4.2, we introduce the general theory involved in chance-constrained programming under distributional ambiguity, the Wasserstein ambiguity set, and the recently-developed optimal transport distance between Gaussian Mixture Models. In Section 4.3, we derive the distributionally robust chance-constrained programming model. In Section 4.4, we discuss the results of applying the derived method to an illustrative example. We further discuss the worst-case expectation distribution associated with this example. In Section 4.5, we apply the derived method to a practical chemical process case study. Finally, we summarize the key findings of our work in Section 4.6.

4.2 Theory

4.2.1 Distributionally robust optimization

Distributionally robust optimization (DRO) is a method of optimization under uncertainty wherein the user seeks to optimize the expected value of a cost function under its worst-case realization over an ambiguity set of probability distributions. The general form of the DRO problem is given as,

$$\min_{x \in X} \max_{\mathbb{P}(\xi) \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\mathcal{L}(x, \xi)] \tag{4.1}$$

Here, $x \in X$ denotes the set of constraints defining a feasible solution space for the decision variables, and ξ denotes the parametric uncertainty involved in the cost function $\mathcal{L}(x, \xi)$. For the metric-based ambiguity sets used in this work, the general form of the ambiguity set $\mathcal{P}$ may be defined as,

$$\mathcal{P} = \{\mathbb{P}(\xi) : \mathcal{M}(\mathbb{P}, \mathbb{P}^0) \leq \epsilon\} \tag{4.2}$$

The membership of the elements of the ambiguity set, hereby termed as candidate distributions ($\mathbb{P}$), is defined using the radius ϵ. The ambiguity set contains all candidate probability distributions $\mathbb{P}$ which are within a certain ϵ-magnitude of a metric $\mathcal{M}$ relative to a nominal distribution $\mathbb{P}^0$. The size of the ambiguity set is controlled via ϵ which further introduces the level of distributional ambiguity considered, and must be carefully tuned to ensure that $\mathcal{P}$ is not unnecessarily large thereby admitting unnecessary distributions for consideration under the worst-case setting. A schematic representation of a metric-based ambiguity set is depicted in Figure 4.1. The nominal distribution $\mathbb{P}^0$ is usually defined on available information of the uncertainty ξ; for

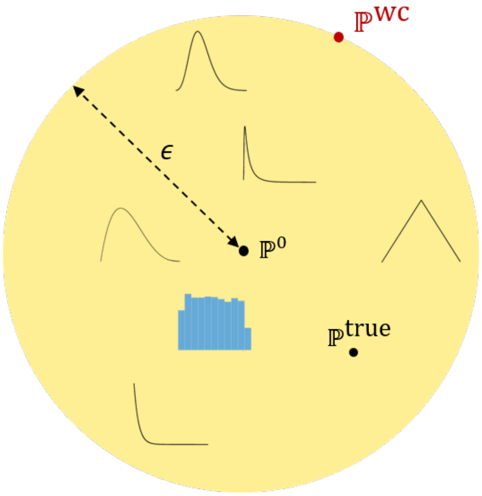


Figure 4.1: A schematic representation of a metric-based ambiguity set ($\mathcal{P}$) used for DRO. $\mathbb{P}^{true}$ is the true underlying probability distribution of the uncertainty (ξ) in the problem, which is not known exactly by the user. $\mathbb{P}^0$ is the nominal distribution estimated from available information on ξ. ϵ is the user-defined radius of $\mathcal{P}$. $\mathbb{P}^{wc}$ refers to that distribution in $\mathcal{P}$ that gives the worst-case expected performance of the problem. The different distributions are among the candidate distributions ($\mathbb{P}$) located within ϵ−magnitude of a metric from $\mathbb{P}^0$

.

instance, it may be estimated as a uniformly weighted empirical distribution using N sampled data realizations of $\xi := \{\xi_1^0, \xi_2^0, ..., \xi_N^0\}$ as,

$$\mathbb{P}^0(\xi_j^0) = \frac{1}{N}, \quad \forall 1 \leq j \leq N \tag{4.3}$$

Under the definition in Model 4.1, the DRO model is well-equipped to deal with optimization problems involving parametric uncertainty in the objective function. When the problem involves uncertainty in the its constraints, chance-constrained programming is a popular stochastic programming method to obtain optimal solutions. This approach is discussed in the next section.

4.2.2 Chance-constrained programming

As mentioned in Section 4.1, stochastic programming methods for optimization under uncertainty utilize the probability distribution information of the underlying uncertainty. Chance-constrained programming (CCP), is one such method in which the uncertainty-riddled constraints in an optimization problem are modeled as probabilistic constraints. More specifically, CCP models the problem such that the optimal solution obtained optimizes the given objective function while simultaneously satisfying the constraint defining its feasibility to a user-defined level. The general form of a CCP problem is given as,

$$\min_{x \in X} \quad f(x) \tag{4.4a}$$

$$\text{s.t.} \quad \Pr\Big(g_i(x,\xi) \leq 0, \quad \forall 1 \leq i \leq m\Big) \geq 1 - \delta \tag{4.4b}$$

Here, $\delta \in [0,1]$ defines the level of "strictness" (also termed as constraint violation allowance) to which the m constraints must be satisfied; it is usually set to a small value such as 0.1 or 0.05 to ensure that the constraint 4.4b is satisfied with a probability of 90% or 95%, respectively. Model 4.4 considers no distributional ambiguity with respect to the distribution of ξ. Under the setting discussed in Section 4.2.1, the distributionally robust chance-constrained programming (DRCCP) problem may be defined as,

$$\min_{x \in X} \quad f(x) \tag{4.5a}$$

$$\text{s.t.} \quad \min_{\mathbb{P} \in \mathcal{P}} \quad \Pr_{\mathbb{P}}\Big(g_i(x,\xi) \leq 0, \quad \forall 1 \leq i \leq m\Big) \geq 1 - \delta \tag{4.5b}$$

Constraint 4.5b restricts the model to find an optimal solution such that the worst-case probability of satisfaction of the feasibility constraints, over all the distributions in $\mathcal{P}$, is at least $1 - \delta$. It must be noted that the CCP models in (4.4) and (4.5) consider the "joint" chance-constrained (JCC) problem which is a setting under which m constraints must be simultaneously satisfied to the required level. When $m = 1$, the problem is termed an "individual" chance-constrained (ICC) problem.

4.2.3 Wasserstein distance

Optimal transport (OT) theory is a mathematical concept that borrows from the logistic optimization problem of matching supply to demand at least cost, in order to address the topic of computing the similarity between probability distributions. Under this view, the OT problem quantifies the effort required to "reshape" one probability distribution to fit another as a metric of similarity between the distributions.

Consider the probability spaces $\zeta \sim \mathbb{P}^{(1)}$ and $\omega \sim \mathbb{P}^{(2)}$, and an associated pairing cost function $\theta(\zeta, \omega)$ that may be described by a distance. The p–Wasserstein distance between $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ is computed as,

$$W_p(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}) := \left(\inf_{\pi \in \Pi(\mathbb{P}^{(1)} \times \mathbb{P}^{(2)})} \int \theta(\zeta, \omega)^p d\pi(\zeta, \omega) \right)^{1/p} \tag{4.6}$$

Here, $\theta(\zeta, \omega)$ is a measurable, nonnegative distance function, usually norm-based, that describes the cost of transporting probability mass from elements of ζ to those of ω. Π denotes the set of all joint distributions whose marginal distributions are $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$. The optimal objective value of the problem is denoted as the "optimal transport (OT) distance", which is a quantitative similarity measure between $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$, while the "optimal transport plan" refers to the least cost method of transporting probability masses from ζ to ω. The 1-Wasserstein (W_1) distance is obtained by setting $p = 1$ in 4.6 as,

$$W_1(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}) := \min_{\pi \in \Pi(\mathbb{P}^{(1)}, \mathbb{P}^{(2)})} \int ||\zeta - \omega|| d\pi(\zeta, \omega) \tag{4.7}$$

In practical applications, the discrete form of the 1-Wasserstein optimal transport problem in 4.7 is used. For the discrete distributions $\tilde{\mathbb{P}}^{(1)}$ and $\tilde{\mathbb{P}}^{(2)}$ supported on A and B samples, respectively, the 1-Wasserstein distance is computed as,

$$W_1(\tilde{\mathbb{P}}^{(1)}, \tilde{\mathbb{P}}^{(2)}) := \min_{\pi_{a,b} \geq 0} \quad \sum_{a=1}^{A} \sum_{b=1}^{B} ||\zeta_a - \omega_b|| \pi_{a,b} \tag{4.8a}$$

$$\text{s.t.} \quad \sum_{b=1}^{B} \pi_{a,b} = \rho_a, \quad \forall 1 \le a \le A \tag{4.8b}$$

$$\sum_{a=1}^{A} \pi_{a,b} = \rho_b, \quad \forall 1 \le b \le B \tag{4.8c}$$

Here, the constraints 4.8b and 4.8c denote the probability mass conservation constraints imposed on the discrete transport problem, while ρ_a and ρ_b denote the probability mass associated with the support elements ζ_a and ω_b of the distributions $\tilde{\mathbb{P}}^{(1)}$ and $\tilde{\mathbb{P}}^{(2)}$, respectively. An illustration of the discrete optimal transport problem is presented in Figure 4.2.

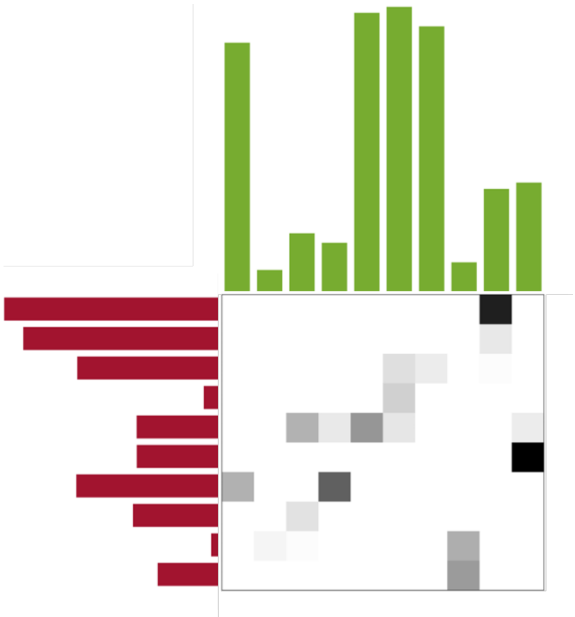


Figure 4.2: A schematic depiction of the discrete optimal transport problem (Model 4.8). The color intensity of the transport map depicts the amount of probability mass transported from the supporting elements of $\tilde{\mathbb{P}}^{(1)}$ (green) to those of $\tilde{\mathbb{P}}^{(2)}$ (red). The length of the bars depict the probability masses of the elements supporting $\tilde{\mathbb{P}}^{(1)}$ and $\tilde{\mathbb{P}}^{(2)}$

.

The 1-Wasserstein distance is a popularly used metric that has been used extensively in literature (Rahimian and Mehrotra 2019) to construct ambiguity sets for

DRCCP. The 1-Wasserstein ambiguity set is defined as,

$$\mathcal{P}_W = \{\mathbb{P} : W_1(\mathbb{P}^0, \mathbb{P}) \le \epsilon_W\} \tag{4.9}$$

That is, the 1-Wasserstein ambiguity set $\mathcal{P}_W$ is defined as the set of all candidate probability distributions $\mathbb{P}$ which are located within a certain ϵ_W-magnitude of 1-Wasserstein distance relative to the nominal distribution $\mathbb{P}^0$. Here, $\mathbb{P}^0$ and $\mathbb{P}$ may be considered analogous to $\tilde{\mathbb{P}}^{(1)}$ and $\tilde{\mathbb{P}}^{(2)}$ in Model 4.8, respectively.

4.2.4 Optimal transport between Gaussian Mixture Models

In this work, we leverage a variant of optimal transport developed by (Chen *et al.* 2018) for Gaussian mixture models (GMMs) to construct an ambiguity set for DRCCP. While the Wasserstein distance applies optimal transport in the Euclidean space, the authors extend this idea to developing optimal transport between GMMs, wherein each GMM is treated as a discrete weighted measure of Gaussian support elements, in a space of Gaussian distributions. Specifically, the authors consider each marginal $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ as a GMM equivalent to an finite-dimensional discrete measure of weights supported by Gaussian component distributions as follows,

$$\mathbb{P}^{(1)} := \sum_{l=1}^{L^{(1)}} w_l^{(1)} \nu_l^{(1)}, \quad \mathbb{P}^{(2)} := \sum_{l=1}^{L^{(2)}} w_l^{(2)} \nu_l^{(2)} \tag{4.10}$$

to solve the following optimal transport problem by solving the following discrete linear programming (LP) problem, henceforth termed the OT-GMM problem in this work,

$$W_d^2(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}) = \min_{\pi \ge 0} \quad \sum_{l=1}^{L^{(1)}} \sum_{l'=1}^{L^{(2)}} c_{l,l'} \pi_{l,l'} \tag{4.11a}$$

$$\text{s.t.} \quad \sum_{l'=1}^{L^{(2)}} \pi_{l,l'} = w_l^{(1)}, \quad \forall 1 \le l \le L^{(1)} \tag{4.11b}$$

$$\sum_{l=1}^{L^{(1)}} \pi_{l,l'} = w_{l'}^{(2)}, \quad \forall 1 \le l' \le L^{(2)} \tag{4.11c}$$

We denote the "optimal transport distance between GMMs" as $W_d(\mathbb{P}^{(1)}, \mathbb{P}^{(2)})$ in our work, which is the square root of the optimal value of the above problem 4.11. It may be noted that the cost $c_{l,l'}$ of transport between the Gaussian components of the GMM marginals may be computed using closed form expression for optimal transport between Gaussian marginals, namely $\zeta \sim \nu^{(1)} : \mathbb{N}(\mu^{(1)}, \Sigma^{(1)})$ and $\omega \sim \nu^{(2)} : \mathbb{N}(\mu^{(2)}, \Sigma^{(2)})$, which is established in literature (Takatsu 2011),

$$W_2(\nu^{(1)}, \nu^{(2)})^2 := \min_{\gamma \in \Gamma(\nu^{(1)}, \nu^{(2)})} \int \|\zeta - \omega\|^2 d\pi(\zeta, \omega) \tag{4.12a}$$

$$= \|\mu^{(1)} - \mu^{(2)}\|^2 + \mathrm{Tr}\left[\Sigma^{(1)} + \Sigma^{(2)} - 2\left(\left[\Sigma^{(1)}\right]^{\frac{1}{2}} \Sigma^{(2)} \left[\Sigma^{(1)}\right]^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \tag{4.12b}$$

$W_d(\mathbb{P}^{(1)}, \mathbb{P}^{(2)})$ defines a metric on the GMM distributional space. It may be noted that (Delon and Desolneux 2020) also propose a Wasserstein-type metric for GMMs; the authors further extend their formulation to probability distributions that are not GMMs by proposing a similarity metric that is a combination of their Wasserstein-type distance and the Kullback-Leibler (KL) Divergence. (Dusson *et al.* 2023) extended their work (Delon and Desolneux 2020) to generic mixture models to provide an optimal transport-type metric to measure the similarity between spaces described by location-scatter atoms, and group-invariant measures. In the following section, we illustrate the use of $W_d(\mathbb{P}^{(1)}, \mathbb{P}^{(2)})$ as an alternative to the 1-Wasserstein metric for DRCCP, and derive a tractable model formulation for the same.

4.3 Proposed distributionally robust chance-constrained optimization method

4.3.1 Preliminaries and background

We start with the general formulation of the DRCCP problem in Model 4.5 wherein the worst-case constraint satisfaction is rewritten from the lens of worst-case constraint violation,

$$\min_{x \in X} \quad f(x) \tag{4.13a}$$

$$\text{s.t.} \quad \max_{\mathbb{P} \in \mathcal{P}} \quad \Pr_{\mathbb{P}}\Big(g_i(x,\xi) > 0, \quad \forall 1 \leq i \leq m\Big) \leq \delta \tag{4.13b}$$

Using an epigraph reformulation of the constraints within the joint chance constraint, Model 4.13 is modified as,

$$\min_{x \in X} \quad f(x) \tag{4.14a}$$

$$\text{s.t.} \quad \max_{\mathbb{P} \in \mathcal{P}} \quad \Pr_{\mathbb{P}}\Big(\max_{1 \leq i \leq m} \quad g_i(x,\xi) > 0\Big) \leq \delta \tag{4.14b}$$

Solving Model 4.14 for its current form of the probabilistic constraint 4.14b poses difficulties for a number of reasons chief of which is that its feasible region may not be convex, or quasi-convex, even for convex functions $g_i(x,\xi), \quad \forall i$ in addition to the need for the computation of multidimensional integrals to check the feasibility of a point. The JCC may be quasi-convex for a number of special cases, such as when $g_i(x,\xi), \quad \forall i$ are all quasi-convex functions of (x,ξ), and ξ has a log-concave distribution (Prékopa 2003). It may be noted that all log-concave distributions are necessarily unimodal (An 1997), and therefore, in practical applications where ξ may follow multimodal distributions, as in our assumption for this work, this condition for quasi-convexity of the JCC, and subsequently, convexity of the JCCP does not hold.

In this context, a better alternative to dealing with the probabilistic constraint would be to use a convex conservative approximation to the same, a number of which have been studied in the literature. One popular approach is to use the Conditional Value-at-Risk (CVaR) approximation (Rockafellar, Uryasev, *et al.* 2000). It may be noted that other approximation techniques such as quadratic approximation (Ben-Tal and Nemirovski 2000), and Bernstein approximation (Nemirovski and Shapiro 2007) are also available. In this work, we use the CVaR approximation to the probabilistic DRCCP.

The probabilistic constraint in 4.14b is first transformed to its indicator function-based form, and the CVaR approximation is applied to the expectation of the indicator function reformulation of constraint 4.14b to give,

$$\min_{x \in X} \quad f(x) \tag{4.15a}$$

$$\text{s.t.} \quad \max_{\mathbb{P} \in \mathcal{P}} \quad \min_{\eta} \quad \eta + \frac{1}{\delta} \mathbb{E}_{\mathbb{P}} \left[\left(\max_{1 \le i \le m} \quad g_i(x, \xi) - \eta \right)^{+} \right] \le 0 \tag{4.15b}$$

Here, $\big(g(x)\big)^{+} := \max\big(g(x), 0\big)$. A detailed derivation of constraint 4.15b from 4.14b may be found in Nemirovski and Shapiro (2007). The "min" and "max" operators in Constraint 4.15b may be rearranged to give,

$$\min_{x \in X} \quad f(x) \tag{4.16a}$$

$$\text{s.t.} \quad \min_{\eta} \quad \eta + \frac{1}{\delta} \underbrace{\max_{\mathbb{P} \in \mathcal{P}} \quad \mathbb{E}_{\mathbb{P}} \left[\left(\max_{1 \le i \le m} \quad g_i(x, \xi) - \eta \right)^{+} \right]}_{\text{Worst-case expectation problem}} \le 0 \tag{4.16b}$$

The worst-case expectation problem indicated in Model 4.16 incorporates the distributional ambiguity considered in this DRCCP formulation. In the following section, we derive the worst-case expectation problem for an ambiguity set constructed using the optimal transport distance between Gaussian mixture models described in Section 4.2.4.

4.3.2 Worst-case expectation problem under the W_d ambiguity set

To address the inner worst-case expectation problem, consider a general form as follows,

$$\max_{\mathbb{P}\in\mathcal{P}} \ \mathbb{E}_{\mathbb{P}}\big[\mathcal{L}(\xi)\big] \tag{4.17}$$

Under the W_d distance, Model 4.17 may be articulated as,

$$\begin{aligned} \max \ \ & \mathbb{E}_{\mathbb{P}}\big[\mathcal{L}(\xi)\big] && \text{(4.18a)} \\ \text{s.t.} \ \ & W_d(\mathbb{P}^0, \mathbb{P}) \leq \epsilon_d && \text{(4.18b)} \end{aligned}$$

where ϵ_d denotes the radius of the ambiguity set defined by the W_d distance. A schematic illustration of the ambiguity set is given in Figure 4.3. When $\mathbb{P}^0$ and $\mathbb{P}$ are modeled as Gaussian mixture models (GMMs), constraint 4.18b may be explicitly written using the optimal transport problem in Model 4.11 as,

$$\begin{aligned} \max \ \ & \mathbb{E}_{\mathbb{P}}\big[\mathcal{L}(\xi)\big] && \text{(4.19a)} \\ \text{s.t.} \ \ & \min_{\pi_{l,l'}\geq 0} \ \sum_{l=1}^{L}\sum_{l'=1}^{L} c_{l,l'}\pi_{l,l'} \leq \epsilon_d^2 && \text{(4.19b)} \\ & \sum_{l'=1}^{L} \pi_{l,l'} = w_l^0, \quad \forall 1 \leq l \leq L && \text{(4.19c)} \\ & \sum_{l=1}^{L} \pi_{l,l'} = w_{l'}, \quad \forall 1 \leq l' \leq L && \text{(4.19d)} \end{aligned}$$

for the nominal ($\mathbb{P}^0$) and candidate ($\mathbb{P}$) distributions modeled as Gaussian mixture models,

$$\begin{aligned} \mathbb{P}^0 &:= w_1^0\nu_1^0 + w_2^0\nu_2^0 + \ldots + w_L^0\nu_L^0 && \text{(4.20a)} \\ \mathbb{P} &:= w_1\nu_1^0 + w_2\nu_2^0 + \ldots + w_L\nu_L^0 && \text{(4.20b)} \end{aligned}$$

and where, for the Gaussian components $\nu_l^0 := \mathbb{N}(\mu_l^0, \Sigma_l^0)$ and $\nu_{l'}^0 := \mathbb{N}(\mu_{l'}^0, \Sigma_{l'}^0)$,

$$c_{l,l'} := W_2(\nu_l^0, \nu_{l'}^0)^2 = ||\mu_l^0 - \mu_{l'}^0||^2 + \mathrm{Tr}\left[\Sigma_l^0 + \Sigma_{l'}^0 - 2\left(\left[\Sigma_l^0\right]^{\frac{1}{2}} \Sigma_{l'}^0 \left[\Sigma_l^0\right]^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \tag{4.21}$$

The ambiguity set is defined here using ϵ_d^2 since the W_d distance is defined as the square root of the objective value of the OT-GMM problem. An important point to note here is that the candidate distributions ($\mathbb{P}$) are defined on the same Gaussian support elements as the nominal distribution ($\mathbb{P}^0$). In doing so, we incorporate some amount of first- and second-order statistical moment information identified from uncertainty data realizations into the defintion of the ambiguity set. The expectation $\mathbb{E}_{\mathbb{P}}\big[\mathcal{L}(\xi)\big]$ with respect to a GMM may be written as $\sum_{l'=1}^{L} w_{l'} \mathbb{E}_{\nu_{l'}^0}\big[\mathcal{L}(\xi)\big]$. Dropping the "min" operator in constraint 4.19b, and rewriting the problem in terms of $\pi_{l,l'}$ only,

$$\max_{\pi_{l,l'}} \quad \sum_{l=1}^{L}\sum_{l'=1}^{L} \pi_{l,l'} \mathbb{E}_{\nu_{l'}^0}\big[\mathcal{L}(\xi)\big] \tag{4.22a}$$

$$\text{s.t.} \quad \sum_{l=1}^{L}\sum_{l'=1}^{L} c_{l,l'}\pi_{l,l'} \leq \epsilon_d^2 \tag{4.22b}$$

$$\sum_{l'=1}^{L} \pi_{l,l'} = w_l^0, \quad \forall 1 \leq l \leq L \tag{4.22c}$$

$$\pi_{l,l'} \geq 0, \quad \forall 1 \leq l \leq L, 1 \leq l' \leq L \tag{4.22d}$$

In order to convert constraint 4.16b into an overall minimization form, the dual of Model 4.22 is taken,

$$\min_{\kappa, y_l} \quad \kappa\epsilon_d^2 + \sum_{l=1}^{L} w_l^0 y_l \tag{4.23a}$$

$$\text{s.t.} \quad y_l \geq -\kappa c_{l,l'} + \mathbb{E}_{\nu_{l'}^0}\big[\mathcal{L}(\xi)\big], \quad \forall 1 \leq l, l' \leq L \tag{4.23b}$$

$$\kappa \geq 0 \tag{4.23c}$$

Strong duality holds due to the linear program nature of the primal problem.

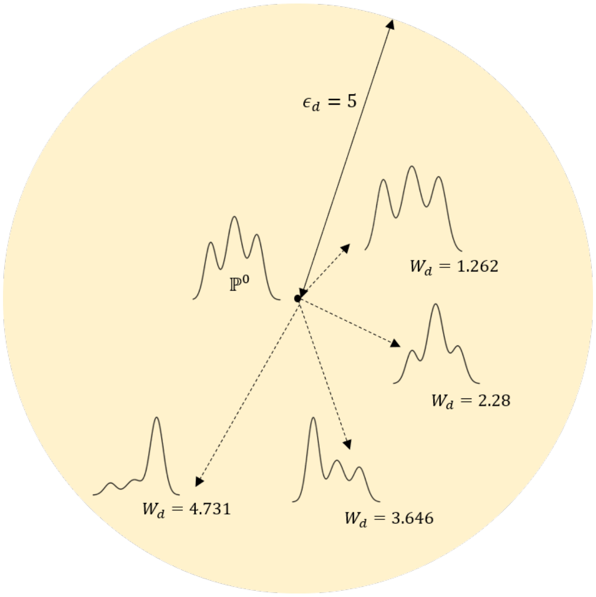


Figure 4.3: An schematic illustration of the W_d distance-based ambiguity set proposed in this work. Here, an ambiguity set of radius $\epsilon_d = 5$ is constructed around the nominal distribution $\mathbb{P}^0$. A number of candidate distributions whose W_d distance from $\mathbb{P}^0$ is at most $\epsilon_d = 5$ are shown.

4.3.3 CVaR-based DRCCP problem under the W_d ambiguity set

Substituting the dual form of the worst-case expectation problem (Model 4.23) into Model 4.16, the following W_d DRCCP problem is obtained,

$$\min_{x \in X} \quad f(x) \tag{4.24a}$$

$$\text{s.t.} \quad \min_{y_l, \eta, \kappa} \quad \eta + \frac{1}{\delta}\Big(\kappa \epsilon_d^2 + \sum_{l=1}^{L} w_l^0 y_l\Big) \leq 0 \tag{4.24b}$$

$$y_l \geq -\kappa c_{l,l'} + \mathbb{E}_{\nu_{l'}^0}\Big[\Big(\max_{1 \leq i \leq m} \; g_i(x, \xi) - \eta\Big)^+\Big], \quad \forall 1 \leq l, l' \leq L \tag{4.24c}$$

$$\kappa \geq 0 \tag{4.24d}$$

The inner minimization operator of 4.24b can be dropped without changing the feasible set. Furthermore, constraint 4.24c requires the computation of the expectation over a max-function imposed on the constraints within the original JCC. To this end, we have chosen to use sample average approximation (SAA) for the inner expectation term as follows,

$$\begin{aligned}
\min_{x\in X, y_l, \eta, \kappa} \quad & f(x) && \text{(4.25a)}\\
\text{s.t.} \quad & \eta + \frac{1}{\delta}\Big(\kappa\epsilon_d^2 + \sum_{l=1}^{L} w_l^0 y_l\Big) \le 0 && \text{(4.25b)}\\
& y_l \ge -\kappa c_{l,l'} + \frac{1}{K}\sum_{k=1}^{K}\Big(\max_{1\le i\le m} \ g_i(x,\xi_k) - \eta\Big)^+, \quad \forall 1 \le l, l' \le L && \text{(4.25c)}\\
& \kappa \ge 0 && \text{(4.25d)}
\end{aligned}$$

Furthermore, an epigraph formulation is used with the introduction of a new variable $t_{k,l'}$ to reformulate $\Big(\max\limits_{1\le i\le m} \ g_i(x,\xi) - \eta\Big)^+$,

$$\begin{aligned}
\min_{x\in X, y_l, \eta, \kappa, t_{j,l'}} \quad & f(x) && \text{(4.26a)}\\
\text{s.t.} \quad & \eta + \frac{1}{\delta}\Big(\kappa\epsilon_d^2 + \sum_{l=1}^{L} w_l^0 y_l\Big) \le 0 && \text{(4.26b)}\\
& y_l \ge -\kappa c_{l,l'} + \frac{1}{K}\sum_{k=1}^{K} t_{k,l'}, \quad \forall 1 \le l, l' \le L && \text{(4.26c)}\\
& t_{k,l'} \ge g_i(x,\xi_k) - \eta, \quad \forall 1 \le k \le K, 1 \le l' \le L, 1 \le i \le m && \text{(4.26d)}\\
& t_{k,l'} \ge 0, \quad \forall 1 \le k \le K, 1 \le l' \le L && \text{(4.26e)}\\
& \kappa \ge 0 && \text{(4.26f)}
\end{aligned}$$

The above model is a deterministic optimization problem, which can be solved using standard optimization solvers depending on the type of the constraint function. For instance, if $f(x)$ and $g_i(x,\xi)$ are all linear functions, then it is a linear programming problem. The above reformulation is applicable for general constraint function $g_i(x,\xi)$. Hence, this method is more general and can be used for various types of optimization problem. This is one advantage compared to the Wasserstein ambiguity set based method, which has limitations in the types of the constraint functions.

4.4 Numerical example

4.4.1 Performance evaluation of the proposed W_d DRCCP method

In this section, we illustrate our DRCCP proposed method (Model 4.26) on a numerical example. This problem (Pagnoncelli *et al.* 2009) refers to a cost minimization problem for a choice of two fertilizers, where x_1 and x_2 refer to the mass amounts of the fertilizers, to achieve certain total nutritional thresholds specified by the contents of the joint chance-constraint. ξ_1 and ξ_2 refer to the uncertain nutritional content in one of two fertilizers, and the objective of this optimal problem is to obtain a decision policy that achieves a minimum nutritional content satisfaction - 12 units for nutrient 1 and 5 units for nutrient 2 - with a user-defined threshold. When the probabilistic constraint is considered under distributional ambiguity, the following distributionally robust joint chance-constrained problem is defined,

$$\min_{x_1,x_2} \quad x_1 + x_2 \tag{4.27a}$$

$$\text{s.t} \quad \min_{P(\xi)\in\mathcal{P}} \quad \mathbb{P}\begin{Bmatrix} \xi_1 x_1 + x_2 \geq 12 \\ \xi_2 x_1 + x_2 \geq 5 \end{Bmatrix} \geq 1 - \delta \tag{4.27b}$$

$$x_1, x_2 \geq 0 \tag{4.27c}$$

We set the joint chance-constraint violation probability to $\delta = 0.05$; that is, we constrained the problem to achieve a minimum joint chance-constraint satisfaction of 95%. As mentioned in the previous sections, our proposed method fits a Gaussian mixture model to the uncertainty data available, to further construct an ambiguity set against which the problem hedges to give an optimal solution. Here, we chose to source the realizations of $[\xi_1, \xi_2]$ from a multimodal process described by the following bivariate 3-component Gaussian mixture model and its corresponding attributes,

$$\mathbb{P}^{\text{true}} := 0.25\mathbb{N}(\mu_1, \Sigma_1) + 0.45\mathbb{N}(\mu_2, \Sigma_2) + 0.3\mathbb{N}(\mu_3, \Sigma_3) \tag{4.28}$$

where,

$$\mu_1 = [1.7, 0.51], \mu_2 = [2.15, 0.72], \mu_3 = [3.4, 0.79],$$
$$\Sigma_1 = \begin{bmatrix} 0.04 & 0 \\ 0 & 0.0006 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 0.06 & 0 \\ 0 & 0.0009 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 0.025 & 0 \\ 0 & 0.0008 \end{bmatrix}$$

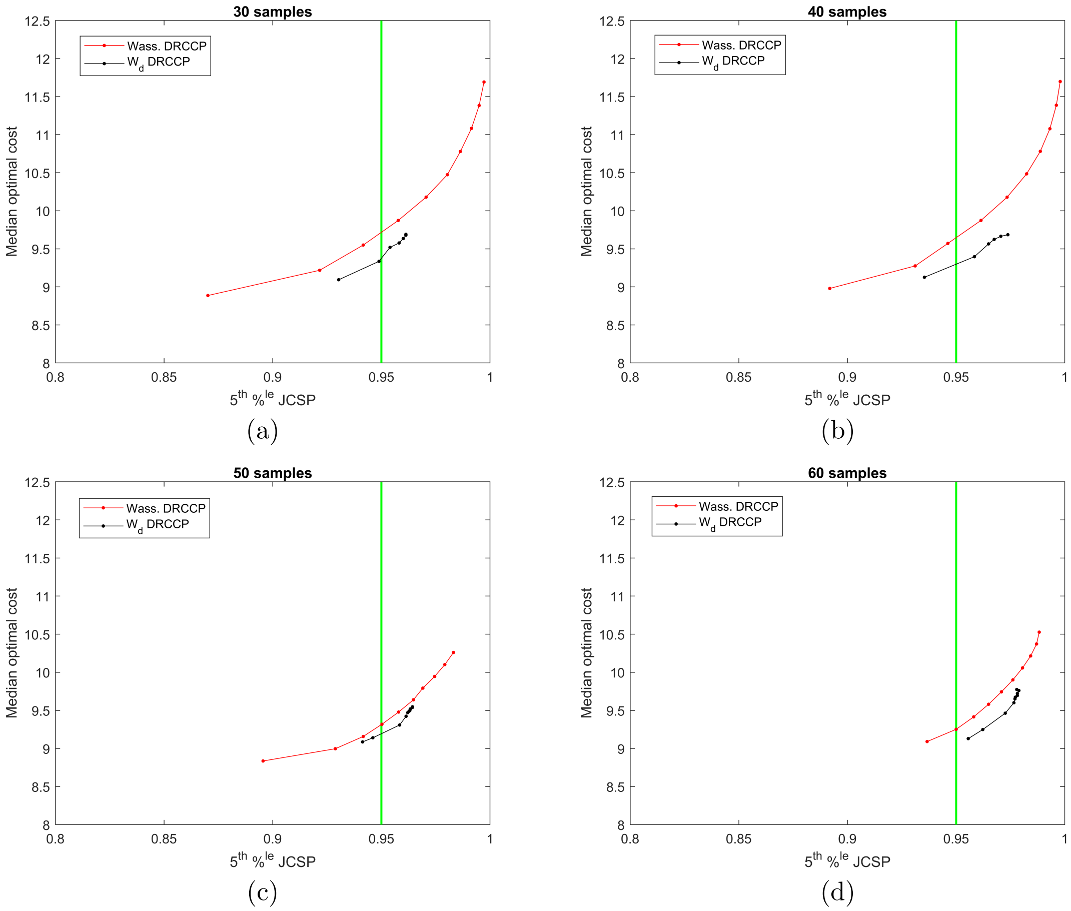
The W_d DRCCP method proposed in this work is based on the Gaussian Mixture Model (GMM) fitted to the available data on the underlying uncertainty in the problem ($\mathbb{P}^0$). In this work, we leverage the *fitgmdist* function available in MATLAB, which uses the expectation maximization (EM) algorithm. The convergence of the EM algorithm to a GMM that describes the multimodal nature of the data well is largely dependent on the choices made regarding the type of covariance matrix for each fitted component, the starting solution for the algorithm, the number of replicates performed, as well as any regularization needed to avoid singularity issues. It may be noted that only attributes of $\mathbb{P}^0$ are utilized in the proposed W_d DRCCP method.

To compare the efficacy of our proposed W_d DRCCP formulation, we ran studies using our method, as well as the established 1-Wasserstein DRCCP method for a test set of uncertainty realizations sourced from $\mathbb{P}^{\text{true}}$. The solution from each DRCCP method is evaluated on the basis of two metrics, namely, the optimal cost associated with the distributionally robust solutions, and the 5th percentile of the joint chance-constraint satisfaction, henceforth denoted as 5th percentile JCSP in this text. The latter metric is a measure of conservatism of the solution. As a general trend in chance-constrained programming, there is a trade-off between the optimal cost associated with a minimization problem and the conservatism of the solution, which is evidenced by the existence of a Pareto front between these metrics. More specifically, a low cost solution displays lower conservatism while a higher cost solution displays more conservatism. In the context of an optimization problem with a minimization objective, it is desirable for this Pareto front to lie in the bottom right hand side of the optimal objective-JCSP plot. It may be noted that larger values of ϵ correspond to a larger degree of ambiguity or a smaller level of confidence regarding the nominal distribution fitted to available data. From a distributionally robust perspective, an increase in the ambiguity set radius is usually associated with an increase in conservatism due to the inclusion of a higher number of probability distributions in the ambiguity set against which the problem hedges. Figure 4.4 illustrates this Pareto front between the median optimal cost over 100 trial datasets, and corresponding the 5^{th} percentile JCSP obtained using a test dataset of 10^6 samples.

Figure 4.4 illustrates the trade-off between the optimal cost and the conservatism of the optimal solution for an increasing range of Wasserstein (ϵ_W) and W_d (ϵ_d) ambiguity set radii, denoted by the red and black curves, respectively, for a choice of 4 different sample sizes $N = \{30, 40, 50, 60\}$. The green line represents the user-defined threshold for a minimum joint chance constraint satisfaction of 95%. From this figure, we show that the trade-off curve pertaining to our proposed method is located lower than that of the Wasserstein method for DRCCP. This result may be explained as follows. In the case of Wasserstein DRCCP, the ambiguity set, being metric-based, includes all probability distributions that are located within a certain ϵ_W-magnitude of Wasserstein distance relative to the nominal distribution. To this end, no further restrictions are placed on the construction of the Wasserstein ambiguity set which, in practice, might include numerous pathological distributions that direct the problem towards finding a more conservative solution, with an associated higher cost, than is required.

However, in the W_d DRCCP approach, the nominal distribution ($\mathbb{P}^0$) is estimated as a GMM from the data realizations available on ξ. Furthermore, the candidate distributions in a W_d ambiguity set are restricted to those GMMs defined on the same Gaussian components (or modes) identified in the data (refer to the definition in 4.20). Therefore, the W_d ambiguity set includes all GMMs based on the components identified in the nominal model ($\mathbb{P}^0$) that are located within a certain ϵ_d-magnitude of W_d distance relative to the nominal distribution. This restriction imposed on the W_d ambiguity set ensures that the W_d DRCCP problem hedges against a smaller number of distributions as compared to the Wasserstein approach; furthermore, this restriction is an informed one. That is, the imposed restriction on the W_d-ambiguity set effectively incorporates first- and second-order statistical moments from the uncertainty data realizations into the model. To that end, the proposed method may be seen as effectively incorporating not only elements of metric-based, but also moment-based ambiguity set construction for DRCCP.

Table 4.1 displays the results of the studies conducted for the numerical example as the median optimal costs and their corresponding tuned ambiguity set radii. We show that for all considered sample sizes (N), as evidenced in Figure 4.4, our proposed W_d DRCCP formulation offers a lower-cost solution compared to the Wasserstein formulation for a tuned 5th percentile JCSP of 95%. It may be noted that for $N = 60$, the W_d DRCCP formulation offers 5th percentile JCSP > 95% even for an ambiguity set radius of $\epsilon_d = 0$. We recall that the CVaR-based approximation of the chance constraint is a convex, *conservative* approximation in order to explain this finding. Finally, we draw the reader's attention to the effect of the number of sampled uncertainty realizations available (N) to fit the nominal distribution. It may be noted that the performance of both Wasserstein and W_d DRCCP methods improves with an increase in N which may be attributed to a better approximated nominal distribution for larger N.



(a)

(b)

(c)

(d)

Figure 4.4: Evolution of the trade-off between median optimal objective and joint chance constraint satisfaction probability (JCSP) for Wasserstein DRCCP (red curve) and the proposed W_d DRCCP (black curve) using (a) 30 samples, (b) 40 samples, (c) 50 samples, and (d) 60 samples of uncertainty realizations to fit $\mathbb{P}^0$. The green line depicts the minimum threshold of 95% JCSP to be achieved by the optimal solution.

Table 4.1: Median optimal objective values and their corresponding tuned ambiguity set radii for 95% JCSP

Sample size	Wasserstein DRCCP	W_d DRCCP
30	9.7143 / 0.0056	9.3604 / 0.2385
40	9.6465 / 0.0050	9.3137 / 0.1700
50	9.3146 / 0.0033	9.1675 / 0.1350
60	9.2537 / 0.0011	9.1292 / 0.0000

4.4.2 Worst-case distributions

In this section, we investigate the worst-case distributions encountered in the Wasserstein and the proposed W_d approaches to DRCCP for the numerical example shown in Model 4.27. As mentioned in Section 4.2.2, DRCCP finds optimal solutions under distributional ambiguity associated with the chance constraint by hedging against the worst-case distribution in the ambiguity set. To this end, it is worthwhile to, first, find the probability distribution associated with the worst-case realization of the probabilistic constraint (henceforth, referred to as $\mathbb{P}^{\text{wc}}$), and then, compare and contrast the worst-case distributions encountered in Wasserstein and W_d DRCCP.

The Wasserstein worst-case expectation distribution is obtained by solving Model 4.17 with the loss function set to $\big(\max\limits_{1\le i\le m} \; g_i(x,\xi) - \eta\big)^+$, and the ambiguity set description taken from Model 4.8. The corresponding Wasserstein worst-case expectation problem is given as,

$$\max_{\pi^W_{j,h}\ge 0} \quad \sum_{j=1}^{N^{(1)}} \sum_{h=1}^{N^{(2)}} \big[\max_{1\le i\le m} \; g_i(x^*_{\text{Wass}}, \xi_h) - \eta^*_{\text{Wass}}\big]^+ \pi^W_{j,h} \tag{4.29a}$$

$$\text{s.t.} \quad \sum_{j=1}^{N^{(1)}} \sum_{h=1}^{N^{(2)}} ||\xi_h - \xi_j|| \pi^W_{h,j} \le \epsilon_W \tag{4.29b}$$

$$\sum_{h=1}^{N^{(2)}} \pi^W_{j,h} = \frac{1}{N}, \quad \forall 1 \le j \le N^{(1)} \tag{4.29c}$$

where $\{\xi_h\}$ denotes the discrete support point for the candidate distributions, while $\{\xi_j\}$ denotes the discrete support of the nominal distribution. $N^{(1)}$ is the number of samples in the nominal empirical distribution, $N^{(2)}$ is the number of support points for the candidate distribution. In this illustration, we choose $N^{(2)} >> N^{(1)}$ to investigate the worst-case distribution. $\pi^W_{j,h}$ denotes the OT plan between the empirical (uniformly-weighted) discrete nominal distribution and the Wasserstein worst-case distribution.

The W_d worst-case expectation distribution is obtained by solving Model 4.22 with the same loss function $\left(\max\limits_{1\le i\le m} \; g_i(x,\xi)-\eta\right)^{+}$

$$\max_{\pi^d_{j,h}\ge 0} \quad \sum_{l=1}^{L}\sum_{l'=1}^{L}\frac{1}{K}\sum_{j=1}^{K}\left[\max_{1\le i\le m} \; g_i(x^*_{W_d},\xi_{j,l'})-\eta^*_{W_d}\right]^{+}\pi^d_{l,l'} \tag{4.30a}$$

$$\text{s.t.} \quad \sum_{l=1}^{L}\sum_{l'=1}^{L} c_{l,l'}\pi^d_{l,l'} \le \epsilon_d^2 \tag{4.30b}$$

$$\sum_{l'=1}^{L}\pi^d_{l,l'} = w^0_l, \quad \forall 1\le l\le L \tag{4.30c}$$

where $\{\xi_{j,l'}\}$ refer to the sampled uncertainty realizations from each Gaussian component (ν^0_l) of the GMM fitted to the data $\{\xi^0_j\}$. $\pi^d_{l,l'}$ denotes the OT plan between the fitted nominal GMM and the W_d worst-case (GMM) distribution. Both models 4.29 and 4.30 are solved using their respective DRCCP model solutions $(x^*_{\text{Wass}}, x^*_{W_d})$ and CVaR approximation variables $(\eta^*_{\text{Wass}}, \eta^*_{W_d})$. The discrete probability masses $\{\rho_h\}$ associated with the Wasserstein candidate support set $\{\xi_h\}$ are computed from $\pi^W_{j,h}$. The Gaussian component weights $w_{l'}$ associated with the W_d candidate Gaussian components are computed from $\pi^d_{l,l'}$. These weights were further used to visualize the worst-case distributions for Wasserstein and W_d DRCCP in Figure 4.5.

Figures 4.5a and 4.5b illustrate the discrete worst-case distribution in 2-dimensions (red bars) for an instance of the problem in 4.27 solved using Wasserstein DRCCP with an ambiguity set radius $\epsilon_W = 0.0022$. Figures 4.5c and 4.5d illustrate the GMM in 2-dimensions (red curves) corresponding to the worst-case distribution for the same instance of Model 4.27 solved using W_d DRCCP with $\epsilon_d = 0.0022$. The blue curve illustrates the true underlying distribution in (4.28).

It may be noted that the discrete worst-case distribution associated with a Wasserstein ambiguity set is supported on at most $N^{(1)}+1$ points (Gao and Kleywegt 2023), which is also verified by the figure above. Therefore, since Wasserstein DRCCP hedges

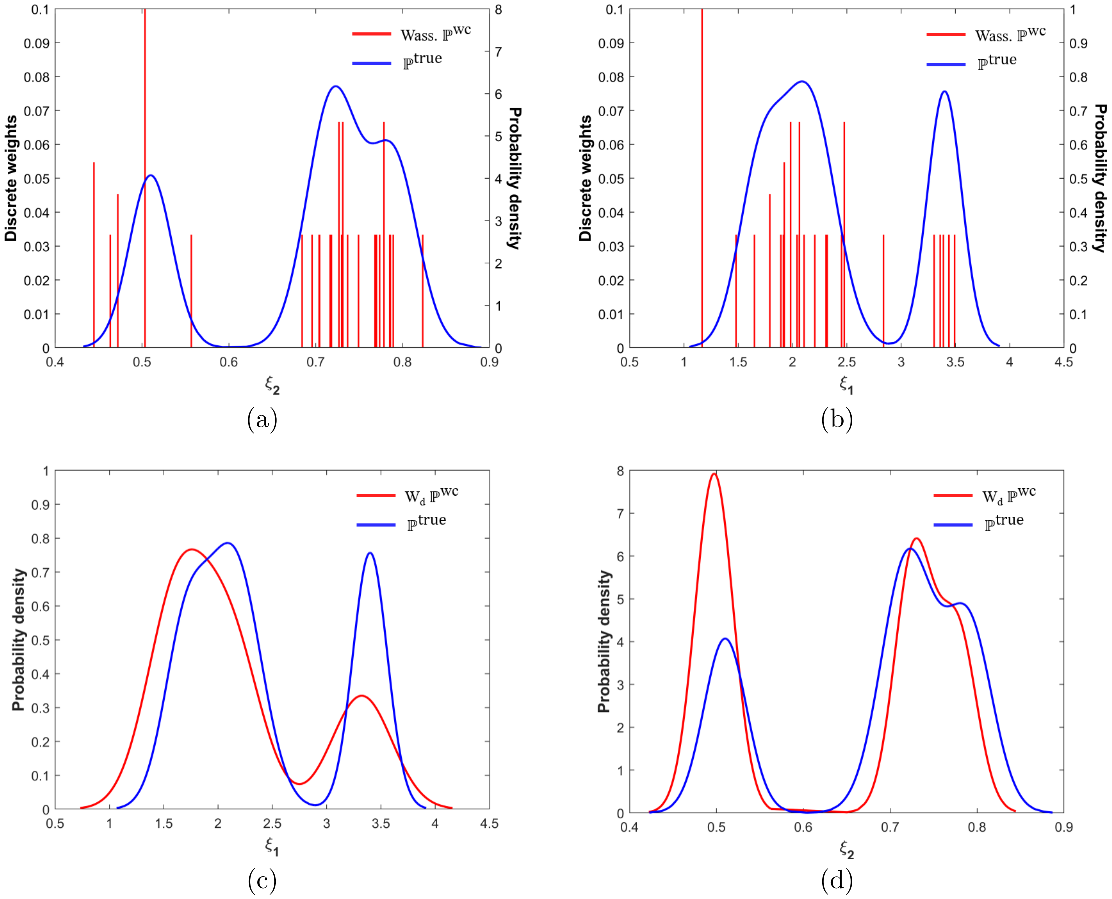


Figure 4.5: An illustration of the worst-case distributions associated with the Wasserstein DRRCP [(a) - (b)] and W_d DRCCP [(c) - (d)] formulations. The red bars in (a) - (b) depict the discrete weights corresponding to the worst-case distribution for the Wasserstein DRCCP approach, while the red curves in (c) - (d) depict the worst-case continuous (GMM) distributions for the W_d DRCCP approach. The blue curves in (a) - (d) depict the true GMM ($\mathbb{P}^{\mathrm{true}}$).

against a discrete distribution, it is not well-equipped to offer optimal (or even feasible in some cases) solutions when the true underlying distribution of uncertainty is continuous in nature. In contrast, the worst-case distribution associated with the W_d ambiguity set used in this work is always a continuous distribution (that is, a GMM). In the case that the true distribution shows multimode feature, the worst-case GMM can well capture this multimodal property. In other words, the proposed method can hedge against the right type of distributions to avoid conservative solution, which is the case by using classical Wasserstein ambiguity set.

4.5 Case study

In this section, we apply the proposed W_d DRCCP formulation to a chemical process design case study under parametric uncertainty. This problem statement was adapted from Example 14.3 (Edgar *et al.* 2001) (along similar lines as in Yang and Li (2023)). The chemical process shown in Figure 4.6 refers to a simplified alkylation process described by Sauer *et al.* (1964) whose total operating profit per day is to be maximized, subject to various performance and economic constraints, and physical relationships.

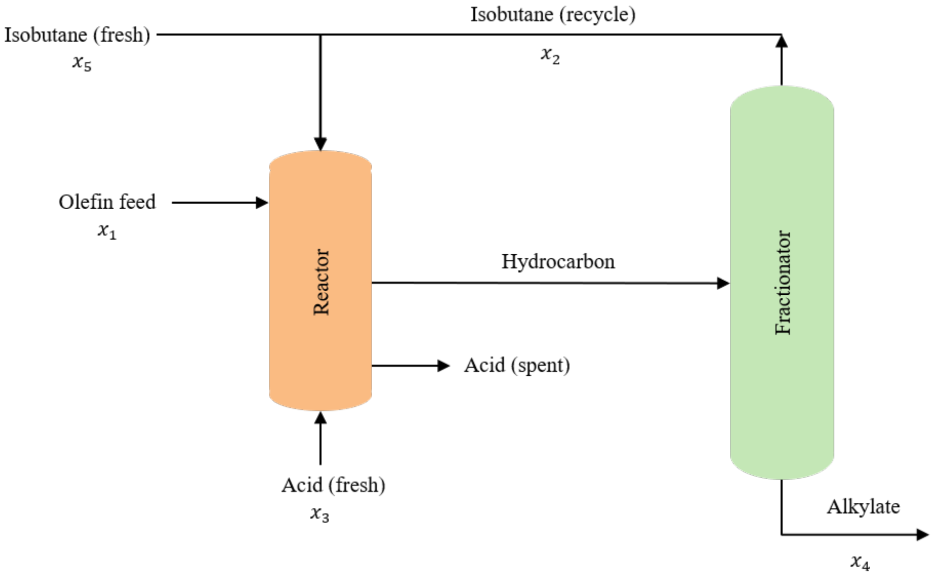


Figure 4.6: A schematic representation of the process flowsheet for acid-catalysed alkylation of olefins.

The objective is to maximize the net profit, calculated from revenue due to alkylate product sales and costs due to olefin feed, isobutane recyle, acid addition, and isobutane makeup streams. The process variables x_p descriptions and bounds are given in Table 4.2. The process model is described through the physical relationships in Equations 4.31. Equation 4.31a describes the volumetric balance over the reactor in order to compute the isobutane makeup flow rate (x_5), accounting for shrinkage.

Equation 4.31b computes the acid strength (by weight percentage) using the acid addition rate (x_3), alkylate yield (x_4), and the acid dilution factor (x_9), assuming that the added acid stream has a strength of 98%. The motor octane number (x_7) is modeled through a nonlinear function of the isobutane-olefin ratio (x_8) and acid strength (x_6) (Equation 4.31c). Equation 4.31d is used to compute the isobutane-olefin ratio (x_8) using the recycle (x_2) and the makeup streams (x_5), and the olefin feed rate (x_1). The acid dilution factor (x_9) is a linear function of the F-4 performance number (x_{10}) (Equation 4.31e).

$$x_5 = 1.22x_4 - x_1 \tag{4.31a}$$

$$x_6 = \frac{98000x_3}{x_4x_9 + 1000x_3} \tag{4.31b}$$

$$x_7 = 86.35 + 1.098x_8 - 0.038x_8^2 + 0.325(x_6 - 89) \tag{4.31c}$$

$$x_8 = \frac{x_2 + x_5}{x_1} \tag{4.31d}$$

$$x_9 = 35.82 - 0.222x_{10} \tag{4.31e}$$

The model also includes two regression models 4.32b and 4.32c for x_4 and x_{10}, respectively. In this work, we relaxed the regression models by introducing user-defined acceptable levels of prediction error. Furthermore, we consider parametric uncertainty in the regression coefficients of the nonlinear model for x_4 by means of a distributionally robust joint chance-constraint imposed on this relationship. The corresponding DRCCP problem is formulated as following,

$$\max_{x} \;\; C_1x_4x_7 - C_2x_1 - C_3x_2 - C_4x_3 - C_5x_5 \tag{4.32a}$$

$$\text{s.t.} \;\; \min_{\mathbb{P}\in\mathcal{P}} \; \Pr_{\mathbb{P}}\begin{pmatrix} x_1(\xi^{(1)} + \xi^{(2)}x_8 + \xi^{(3)}x_8^2) \geq 0.89x_4 \\ x_1(\xi^{(1)} + \xi^{(2)}x_8 + \xi^{(3)}x_8^2) \leq 1.12x_4 \end{pmatrix} \geq 1 - \delta \tag{4.32b}$$

$$-0.89x_{10} \leq 133 + 3x_7 \leq 1.12x_{10} \tag{4.32c}$$

$$\text{Eqs. 4.31a - 4.31e} \tag{4.32d}$$

Table 4.2: Description, bounds and values associated with the operating variables and parameters associated with the alkylation process case study

Variable	Description	Units	Lower bound	Upper bound
x_1	Olefin feed rate	barrels/day	0	2000
x_2	Isobutane recycle rate	barrels/day	0	16000
x_3	Acid addition rate	10^3 lbs/day	0	120
x_4	Alkylate yield rate	barrels/day	0	5000
x_5	Isobutane makeup rate	barrels/day	0	2000
x_6	Acid strength (weight percentage)	-	85	93
x_7	Motor octane number	-	90	95
x_8	Isobutane-olefin ratio	-	3	12
x_9	Acid dilution factor	-	0.01	4
x_{10}	F-4 performance number	-	145	162
Parameter	Description	Units	Value	
c_1	Alkylate product value	\$/octane-barrel	0.063	
c_2	Olefin feed cost	\$/barrel	5.04	
c_3	Isobutane recycle cost	\$/barrel	0.035	
c_4	Acid addition cost	\$/10^3 lbs	10	
c_5	Isobutane makeup cost	\$/barrel	3.36	

We consider multimodal uncertainty sourced from a 3-component GMM.

$$\mathbb{P}^{\text{true}} := 0.25\mathbb{N}(\mu_1, \Sigma_1) + 0.5\mathbb{N}(\mu_2, \Sigma_2) + 0.25\mathbb{N}(\mu_3, \Sigma_3) \tag{4.33}$$

where,

$$\mu_1 = [1.1, 0.099, 0.0058], \mu_2 = [1.12, 0.1317, 0.0067], \mu_3 = [1.14, 0.1670, 0.0071],$$

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 4.9 \times 10^{-5} & 0 & 0 \\ 0 & 8.4 \times 10^{-6} & \\ 0 & 0 & 1.8 \times 10^{-8} \end{bmatrix}$$

Model 4.34 illustrates the W_d DRCCP formulation of the case study.

$$\max_{x,\kappa,t_{k,l'},\eta,y_l} \quad C_1 x_4 x_7 - C_2 x_1 - C_3 x_2 - C_4 x_3 - C_5 x_5 \tag{4.34a}$$

$$\text{s.t.} \quad \eta + \frac{1}{\delta}\left(\kappa \epsilon_d^2 + \sum_{l=1}^{L} y_l\right) \leq 0 \tag{4.34b}$$

$$y_l \geq \left(\frac{1}{K} \sum_{j=1}^{K} t_{k,l'} \right) - \kappa c_{l,l'}, \quad 1 \leq l, l' \leq L \tag{4.34c}$$

$$t_{k,l'} \geq -x_1(\xi_{j,l'}^{(1)} + \xi_{j,l'}^{(2)} x_8 - \xi_{j,l'}^{(3)} x_8^2) + 0.89 x_4, \quad \forall 1 \leq k \leq K, 1 \leq l' \leq L \tag{4.34d}$$

$$t_{k,l'} \geq x_1(\xi_{j,l'}^{(1)} + \xi_{j,l'}^{(2)} x_8 - \xi_{j,l'}^{(3)} x_8^2) - 1.12 x_4, \quad \forall 1 \leq k \leq K, 1 \leq l' \leq L \tag{4.34e}$$

$$133 - 3x_7 + 0.89 x_{10} \leq 0 \tag{4.34f}$$

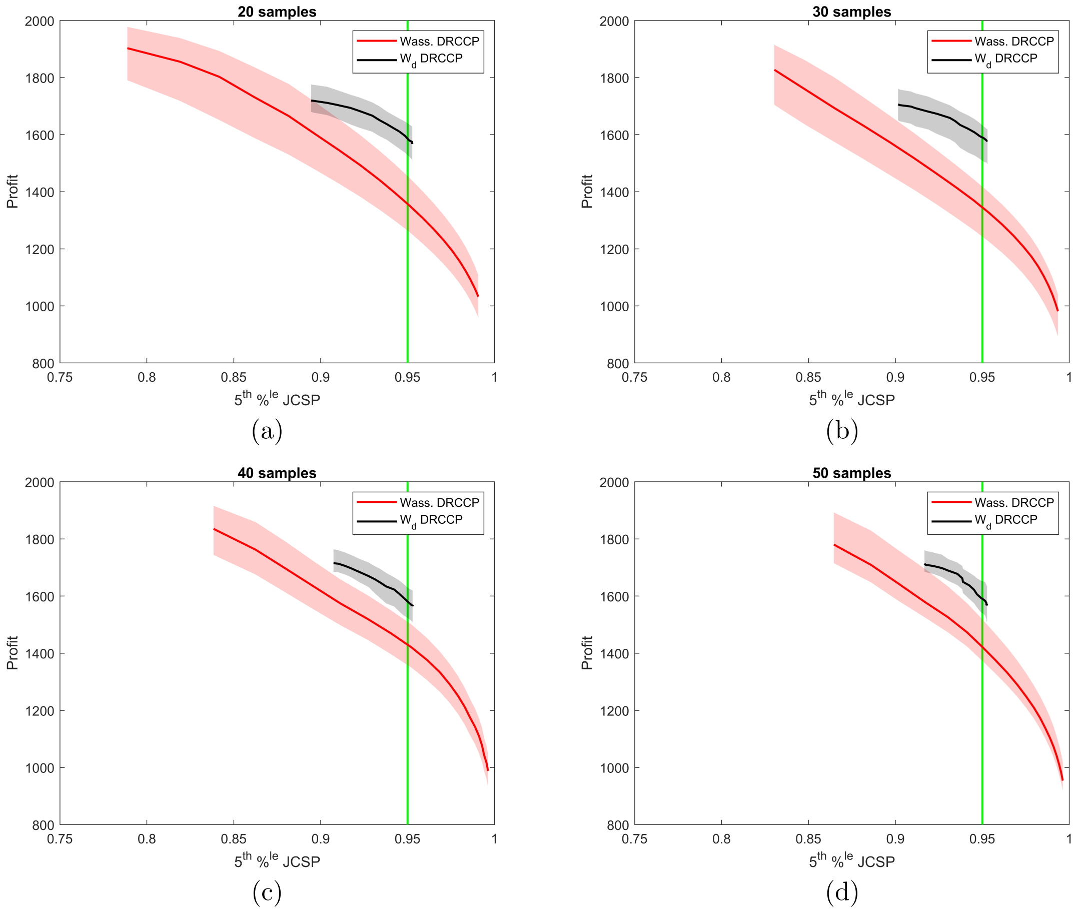
$$-133 + 3x_7 - 1.12 x_{10} \leq 0 \tag{4.34g}$$

$$\kappa \geq 0, t_{k,l'} \geq 0, \quad \forall 1 \leq k \leq K, 1 \leq l' \leq L \tag{4.34h}$$

$$\text{Eqs. 4.31a - 4.31e} \tag{4.34i}$$

The resulting problem in Model 4.34 is a nonlinear programming (NLP) problem in x. In order to compare the practical performance of the proposed W_d formulation, we also solved the DRCCP formulation of this problem using the Wasserstein ambiguity set (also an NLP problem). As in the case of the numerical example, we solved this problem for different sample sizes $N = \{20, 30, 40, 50\}$ for 100 replicate datasets, and tracked the evolution of the optimal objective value and the joint chance constraint satisfaction for increasing ambiguity set radii for $\delta = 0.05$. Figures 4.7 contrast the trade-off between the median optimal profit obtained through the solutions of the proposed W_d DRCCP approach (black curve) and the established Wasserstein DRCCP approach (red curve) and the associated JCSP, wherein the solid lines represent the median profit values and the shaded regions depict the 20th and 80th percentile values. As in the case of the numerical example, it is evident that the proposed W_d DRCCP method is able to leverage the inclusion of statistical moment information to produce solutions with higher profit as compared to the Wasserstein DRCCP, for all levels of JCSP ≥ 95%, based on the relatively higher position of the W_d DRCCP tradeoff curve/Pareto front compared to that of Wasserstein DRCCP. Additionally, for the threshold JCSP of 95%, we observe that the W_d DRCCP method offers optimal profits with a lower variability over repeat trials than the Wasserstein DRCCP method.

In addition to the evident improvement in the optimal objective value, we draw the reader's attention to the slope of the trade-off curves in Figure 4.7 that signifies the rate at which the problems tends to higher levels of JCSP at the cost of lower profits (i.e., conservative solutions). In the case of Wasserstein DRCCP, this slope is seen to be significant for small changes in the ambiguity set radius and therefore, careful tuning of the ϵ_W, usually through cross validation techniques, is required to obtain acceptable levels of conservatism. However, for W_d DRCCP, the optimal solution tends to saturate at higher ϵ_W thus preventing the model from returning solutions with significantly lower profits for JCSP $\geq 95\%$. This phenomenon occurs as a result of the worst-case distribution for W_d DRCCP tending towards a single component in the fitted GMM, and not further changing beyond a certain ϵ_d. This property also naturally leads to a method to compute an acceptable upper bound on ϵ_d from the fitted GMM $\mathbb{P}^0$ with no need for cross validation as in the case of Wasserstein DRCCP. Therefore, the proposed W_d DRCCP may be more user-friendly in terms of tuning of the ambiguity set size/radius than the Wasserstein DRCCP method.



(a) (b) (c) (d)

Figure 4.7: Evolution of the trade-off between optimal objective (profit) and the 5th percentile joint chance constraint satisfaction probability (JCSP) for increasing ambiguity set radii of $\epsilon_W = [0, 0.0001]$ for Wasserstein DRCCP (red), and $\epsilon_d = [0, 0.04]$ for W_d DRCCP (black) using (a) 20 samples, (b) 30 samples, (c) 40 samples, and (d) 50 samples of uncertainty realizations to fit $\mathbb{P}^0$. The solid lines represent the median optimal profits, while the shaded areas represent the region between the 20th and 80th percentile optimal profits for Wasserstein DRCCP (red) and W_d DRCCP (black). The green line depicts the minimum threshold of 95% JCSP to be achieved by the optimal solution.

4.6 Summary

In this work, we propose a novel formulation for distributionally robust optimization in the chance-constrained programming setting (DRCCP) that utilizes a metric-based ambiguity set constructed using the optimal transport distance between supports modeled as Gaussian mixture models (GMMs). A driving force behind this formulation was assessing the effect of incorporating statistical moment information, namely the mean and variance, into the construction of the ambiguity set from the perspective of conservatism of the solution given an acceptable threshold. In our approach, we address this by modeling the available information of the underlying uncertainty, often multimodal in practical applications, as a Gaussian mixture model which is further used as a nominal distribution around which the ambiguity set is built. We illustrated the applicability of our proposed DRCCP formulation on a linear programming numerical example, and developed a formulation of the worst-case distribution problem. We further implemented our formulation on a practical nonlinear chemical process optimization case study, and demonstrated the performance in terms of conservatism, as well as variability. From our results, we show that our formulation provides better optimal solutions with lower levels of conservatism. Furthermore, we show that our method offers solutions that are less sensitive to the ambiguity set size, and are thus less sensitive to tuning of the set radius. Additionally, the W_d DRCCP method proposed in this work is more generalizable than Wasserstein DRCCP as it can be applied to both affine as well as non-affine functions of the primitive uncertainty in the joint chance-constraint. As a future work, the proposed method can be extended to more general mixture models to address the multi-mode nature and possibly outlier data simultaneously.

Chapters 2, 3 and 4 of this thesis utilize variants of the optimal transport problem to address challenges in the mathematical optimization framework. Specifically, Chapter 2's methodologies provide a computationally efficient way to improve the performance of stochastic programming, while the formulations presented in Chapters 3 and 4 offer ways contribute to literature on ways to incorporate multimodal characteristics of data into the ambiguity construction step for distributionally robust optimization. In the following chapters, we pivot to a different use of optimal transport, namely, to address process monitoring challenges. In Chapter 5, we provide a framework to use optimal transport distance as a metric for fault detection. In Chapter 6, we consider the process monitoring problem under uncertainty accounting for ambiguity in the probability distribution describing a multimodal process; here, we utilize the formulation presented in Chapter 3 to provide a study of the worst-case expected performance of a fault detection system.

Chapter 5

Change Point and Fault Detection using the Optimal Transport distance

Abstract: The automation of real-time process monitoring in the industry is an ongoing challenge. Chief among the objectives of monitoring are change point detection and the detection of faults in process variables and sensor measurements. In this chapter, we propose a novel algorithm for change point and fault detection using Kantorovich Distance (KD), a metric induced from optimal transport theory. To evaluate the performance of the proposed method, we first evaluate the change point detection capability of the KD metric for data sampled from various probability distributions. Next, the fault detection performance of the KD metric is evaluated for three cases of faults – sustained bias, drift, and multiple intermittent biases – and contrasted against that of the traditional PCA-based metrics, Q and T^2 statistics. The algorithm is tested on several case studies including a synthetic data, a simulated continuous stirred tank heater system and the benchmark Tennessee Eastman process. The results obtained showcase the superiority of the proposed algorithm over the conventional scheme.

5.1 Introduction

The advent of automated systems in process industries in the past decades has revolutionized process modelling and control to a large extent. In addition to this, process monitoring plays a vital role in ensuring that the process functions at the chosen optimal condition. Process monitoring is routinely carried out in industries for a variety of reasons; an important one among those is abnormal event management (AEM). AEM comprises the following steps: detection of process changes or abnormalities, diagnosis of their origins, and synthesis of control decisions to mitigate them (Venkatasubramanian *et al.* 2003c). This exercise is still heavily dependent upon human supervision and interpretation, due to the broad scope of process abnormalities and their varied signatures. Additionally, it is costly to depend on human supervision for monitoring a process plant with large number of variables, as evidenced by various incidences in history. Hence, there is a need to develop automated process monitoring systems that are capable of dealing with large volumes of process data quickly.

One common monitoring objective in process industries is change point detection. The detection of abrupt process changes is of particular interest in chemical process industries, which operate at certain set optimal conditions deviation from which could cause serious product quality deterioration or process safety violations. The change point detection problem can be solved in two ways: supervised learning methods and unsupervised methods, as briefly detailed by Aminikhanghahi and Cook (2017). Supervised methods of change point detection use a variety of classifiers, such as Support Vector Machines (SVM), Gaussian Mixture Models (GMM) and logistic regression. Desobry *et al.* (2005) used single-class SVM to find dissimilarities between two descriptors of a signal, for abrupt change detection. Han *et al.* (2012) used GMM as a classification technique, for context recognition applications, which is essentially change point detection, using sensor data from a smartphone. Unsupervised methods use techniques such as clustering, likelihood ratios, and probabilistic methods. Kawahara and Sugiyama (2009) proposed the idea of estimation of probability densities for the purpose of change point detection. Kuncheva (2011) proposed a change detection algorithm using a semi-parametric log-likelihood criterion (SPLL). For practical implementation, it is deemed necessary for a change point detection algorithm to be capable of on-line monitoring and have as little detection delay as possible.

Fault detection methods are broadly classified into quantitative model-based, qualitative model-based, and process history-based methods. Both quantitative and qualitative model-based methods require knowledge of the process to be monitored, often relying on first principles. Some commonly used quantitative model-based approaches are Kalman filters and parity relations (Venkatasubramanian *et al.* 2003c). Qualitative model-based methods include the use of fault trees and digraphs (Venkatasubramanian *et al.* 2003a). Process history-based methods, however, only require large amounts of process history data, which is further transformed and fed into the detection and diagnostic algorithms. These methods, also known as data-driven feature extraction methods, are further classified into qualitative and quantitative methods. Some important methods of qualitative feature extraction include expert systems and trend analysis (Venkatasubramanian *et al.* 2003a). Quantitative feature extraction methods are of two types: statistical and non-statistical (Venkatasubramanian *et al.* 2003b). Chief among the non-statistical classifiers are neural networks. Statistical feature extraction methods include Principal Component Analysis (PCA), Partial Least Squares (PLS) and pattern classifiers. Both PCA and PLS are multivariate statistical tools that uncover the underlying trends in high-dimensional correlated data and project them to an alternate space that best depicts the trends. PCA and PLS, both conceptually similar, have been used extensively as modelling techniques for anomaly detection in large volume noisy datasets.

Conventionally, PCA modelling of data is used for fault detection in conjunction with the Q and Hotelling's T^2 statistics, which explain the variance not captured, and captured by the model, respectively, as illustrated by Garcia-Alvarez *et al.* (2009). Numerous variants of PCA have been developed for different types of datasets. An extension of PCA, called dynamic PCA, has been developed for time series process data which exhibits auto-correlation, and its application in fault detection in a flow control valve has been evaluated by Mina and Verde (2005). Another extension of PCA, called multiscale PCA, has been developed by Bakshi (1998) to deal with data containing events whose behavior changes over both time and frequency scales. This approach utilizes the ability of wavelet analysis to extract the deterministic features out of multiscale data, combined with the de-correlation ability of PCA, to detect anomalies in data. Most industrial data are highly non-linear in nature. Therefore, a variant called kernel PCA (KPCA) was developed to deal with data exhibiting non-linear relationships. Samuel and Cao (2014) have shown the fault detection performance of KPCA on the Tennessee Eastman process simulation. In this approach, the non-linear data is transformed into a higher-dimensional space first, and PCA is performed on this transformed dataset. Zhang and Jia (2017) proposed a method called kernel uncorrelated component analysis (KUCA) for complex process monitoring, and applied it to a waste liquor treatment process; the efficacy of this process was highlighted particularly for non-Gaussian and non-linear processes. In addition to refining PCA for various types of datasets, the fault detection capability of various probabilistic and statistical metrics has also been documented in literature. One such metric is the Kullback-Leibler Divergence (KLD), which is a measure of dissimilarity between two probability distributions. Harrou *et al.* (2016) show the superiority of fault detection performance of the KLD metric applied to PLS model residuals, over that of the Q and T^2 statistics. Another metric whose fault detection performance has been evaluated, is the Hellinger Distance. Harrou *et al.* (2017a) have used the Hellinger Distance metric with non-linear projection to latent structures (NLPLS) modelling for fault detection of data from a simulated plug flow reactor.

In this chapter, we propose the use of the Kantorovich Distance (KD), a metric from optimal transport theory, for the purpose of change point and fault detection. In this novel algorithm, we propose to employ KD between PCA model residuals of training and testing data. The chapter also highlights the change detection capabilities of the KD metric and showcases its ability to deal with data from various distributions.

The chapter is organized in the following manner. Section 5.2 explains the theory behind Kantorovich Distance and its various formulations. Section 5.3 studies the change detection, and fault detection capability of KD for time series data; it also presents a contrast between the performance of the KD metric calculated using the linear programming and closed-form approaches for data from different distributions. The novel PCA model-based fault detection scheme is presented in Section 5.4, and its performance is evaluated using two case studies and a benchmark setup, as described in Section 5.5. The conclusions of this work are presented in Section 5.6.

5.2 Optimal Transport Theory and Kantorovich Distance

The concept of Kantorovich Distance (KD) originates from the optimal transport problem, which is the basis of transport-based techniques for data analysis. In the optimal transport problem, the objective is to find the most efficient way of transforming one distribution of mass to another, relative to a given cost function. In recent literature, these techniques are receiving special attention in the field of signal processing due to their ability to compare signals and data from different sources. Kantorovich Distance (KD) is a metric that quantifies the minimum cost needed to redistribute the probability mass between two distributions.

5.2.1 Optimal transport

Monge (1781) initially studied the optimal transport problem. His formulation uses a continuous transport map to assign the spatial correspondence between two distributions $\mathbb{P}$ and $\mathbb{Q}$, with supports X and Y, respectively,

$$M(\mathbb{P},\mathbb{Q}) := \inf_{f\in\mathcal{M}} \left\{ \int_X c\big(x, f(x)\big) \;\; d\mathbb{P}(X) \right\} \tag{5.1}$$

where f is the mapping function to be optimized, c is a given cost function, MP is the set of measure-preserving mappings,

$$\mathcal{M} = \left\{ f : X \to Y \middle| \int_{f^{-1}(A)} d\mathbb{P}(x) = \int_A d\mathbb{Q}(y) \right\} \tag{5.2}$$

Kantorovich (1942) proposed a relaxed formulation, which uses a mass transport plan instead. The main difference to Monge's formulation is that mass splitting is allowed in the transport plan, whereas it is not allowed in the transport map. Kantorovich's formulation is given in Equation 5.3,

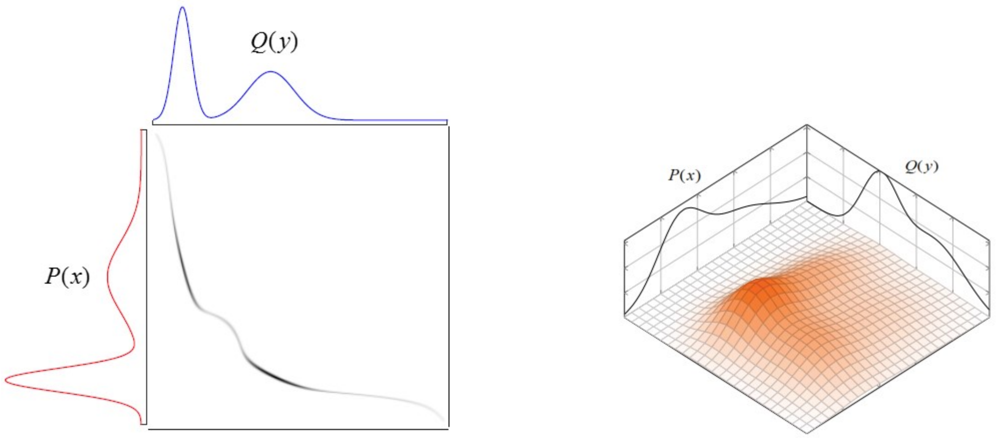
$$K(\mathbb{P},\mathbb{Q}) := \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \left\{ \int_{X\times Y} c\big(x, y\big) \;\; d\gamma(x,y) \right\} \tag{5.3}$$

where $\Gamma(\mathbb{P},\mathbb{Q})$ is the set of all joint distributions with marginal distributions $\mathbb{P}$ and $\mathbb{Q}$,

$$\Gamma = \left\{ \gamma : \gamma(A,Y) = \mathbb{P}(A), \gamma(X,B) = \mathbb{Q}(B) \right\} \tag{5.4}$$

The minimizer γ^* is the optimal transport plan, also called the optimal coupling. Kantorovich's formulation also covers the discrete mass distribution and is more general than Monge's formulation.

To illustrate the above concepts, consider two univariate distributions. Assume that the transport is from the 'source' distribution $\mathbb{P}$ to the 'destination' distribution $\mathbb{Q}$. Monge's transport map seeks a one-to-one map between two points of the two distributions, while Kantorovich's formulation seeks a transport plan which can map a point from the source distribution to multiple points in the destination distribution. Figure 5.1 shows the difference between the two formulations. It is to be noted that the transport map/plan shown in the figure represents only a feasible solution, not necessarily an optimal transport solution.



(a) (b)

Figure 5.1: (a) Representation of the transport map between two univariate distributions (darkness of the curve is proportional to the weight of mass transport), (b) Transport plan between two univariate distributions (the surface plot reflects the difference in probability mass transport)

5.2.2 KD between continuous distributions

The KD metric is a distance function defined between two probability measures with respect to a given cost function. A commonly used method to define the cost function is based on norm,

$$c(x, y) = ||x - y||^p \tag{5.5}$$

For example, the norm can be selected as $||x - y||_1 = \sum_{i=1}^{k} |x_i - y_i|$, $||x - y||_2 = \sum_{i=1}^{k}(|x_i - y_i|)^2$, or other valid forms. With the above norm-based cost function, the corresponding KD between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ is also known as the p-Wasserstein distance for $p \geq 1$,

$$W_p(\mathbb{P}, \mathbb{Q}) = \Big(\inf_{\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})} \int ||x - y||^p d\gamma(x, y) \Big)^{\frac{1}{p}} \tag{5.6}$$

For the special case of $p = 1$, the distance is also known as the Monge-Rubinstein metric or the Earth Mover's distance.

For Gaussian distributions, the closed form expression of the 2-Wasserstein distance has been evaluated by Takatsu (2011). Assume n-dimensional random variable x and y belong to Gaussian measures with mean vectors μ_1 and μ_2, and covariance matrices Σ_1 and Σ_2 respectively, and 2-norm is used for $||x - y||$. Then, the KD between x and y is given by the following closed-form expression in Equation 5.7,

$$W_2(\mathbb{P}, \mathbb{Q}) = \Big\{||\mu_1 - \mu_2||_2^2 + \mathrm{Tr}\Big(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}\Big)\Big\}^{\frac{1}{2}} \tag{5.7}$$

The above conclusion on Wasserstein distance for Gaussian distribution was generalized to elliptically symmetric distributions by Gelbrich (1990). He showed that the formula holds for any two distributions $\mathbb{P}$ and $\mathbb{Q}$ which are translations of distributions whose covariance matrices are related in a certain way. While this condition is fulfilled as long as they are in the same class of elliptically symmetric distributions (Rippl *et al.* 2016). As an example, if both $\mathbb{P}$ and $\mathbb{Q}$ follow t-distributions, the above conclusion still holds.

5.2.3 KD between discrete distributions

In many realistic change detection problems, the information available are discrete data. In such case, the KD can be evaluated through the solution of simple linear optimization problem. The Kantorovich formulation of the optimal transport problem treats the problem as a supply-demand linear programming problem between the elements of two discrete distributions. Consider two discrete distributions given as follows, $X = \sum_{i=1}^{n} p_i \delta_{x_i}, Y = \sum_{j=1}^{m} q_j \delta_{y_j}$.

The KD between two discrete distributions is defined by the optimal objective of the following linear programming (LP) problem,

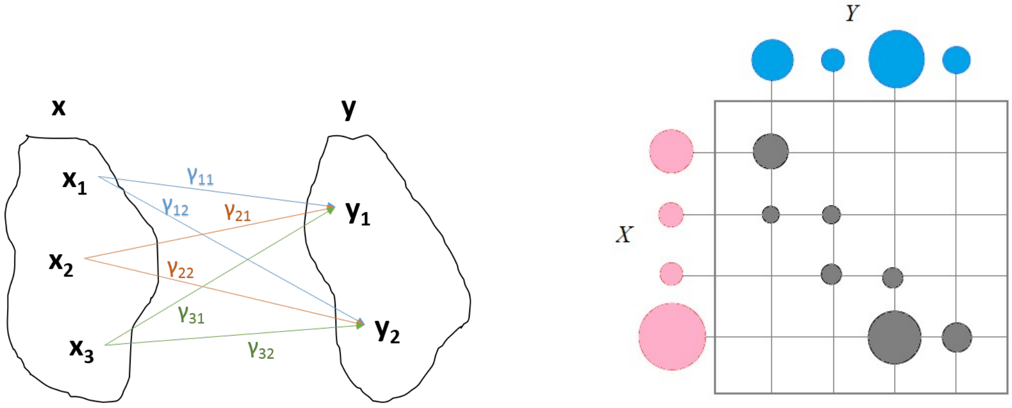
$$\min_{\gamma} \quad \sum_{i=1}^{n}\sum_{j=1}^{m} \gamma_{i,j} c_{i,j} \tag{5.8a}$$

$$\text{s.t.} \quad \sum_{j=1}^{m} \gamma_{i,j} = p_i, \quad \forall i = 1, ..., n \tag{5.8b}$$

$$\sum_{i=1}^{n} \gamma_{i,j} = q_j, \quad \forall j = 1, ..., m \tag{5.8c}$$

$$\gamma_{i,j} \geq 0, \quad \forall i = 1, ..., n, \quad j = 1, ..., m \tag{5.8d}$$

Here, $\gamma_{i,j}$ refers to the amount of probability mass transferred between x_i and y_j, $c(x_i, y_j)$ refers to the 'cost' associated with each transfer of probability mass, p_i and q_j are the probability masses associated with each element in X and Y respectively. The first constraint of Model 5.8 ensures that the probability mass of each x_i is conserved when it is distributed among the elements of Y while the second constraint ensures that the total transport of probability mass from X to each element in Y is exactly equal to the probability mass of each y_j. It is to be noted that the transfer of probability mass from X to Y is not a one-one mapping, but a one-many mapping, as illustrated in Figure 5.2.



(a) (b)

Figure 5.2: Mass transport between two distributions (a) Schematic representation of mass transport, (b) An example of optimal mass transport between discrete distributions

Figure 5.2b illustrates an optimal transport between two discrete distributions.

The circle size is proportional to the probability mass of each element of the two distributions. The grey circle reflects the transportation plan, where the circle size is also proportional to the amount of probability mass transformed.

When the distributional information of the data is not available and the closed form KD formula is not applicable, the LP formulation of KD can be used since it is not dependent on specific distribution.

5.3 Change point detection using KD

Based on the introduction of KD in the previous section, we present the change point detection method for time series data.

5.3.1 KD evaluation for time series data

As described in Section 5.2, the Kantorovich Distance between two probability distributions is the optimal "distance" to be traversed, or rather, the optimal cost incurred, when the probability masses of elements of one distribution are mapped to those of the other distribution. When it is applied to time series data for change detection, a natural way to construct the distribution is to consider each data observation as an element. However, it is observed that grouping individual observations of the data into "segments" and evaluating the KD based on the probability mass mapping of these segments, offers a smoother time evolution of KD. The segmentation strategy employed is illustrated in Figure 5.3. The two signals are split into segments by sliding a moving window of fixed size k samples across the data. Notice that in the extreme case with $k = 1$, it is reduced to the case that each observation is an element of the distribution.

For univariate data, each element of the distribution is a vector and the cost function $c(x_i, y_j) = \|x_i - y_j\|$ can be calculated using vector norm. For multivariate data,

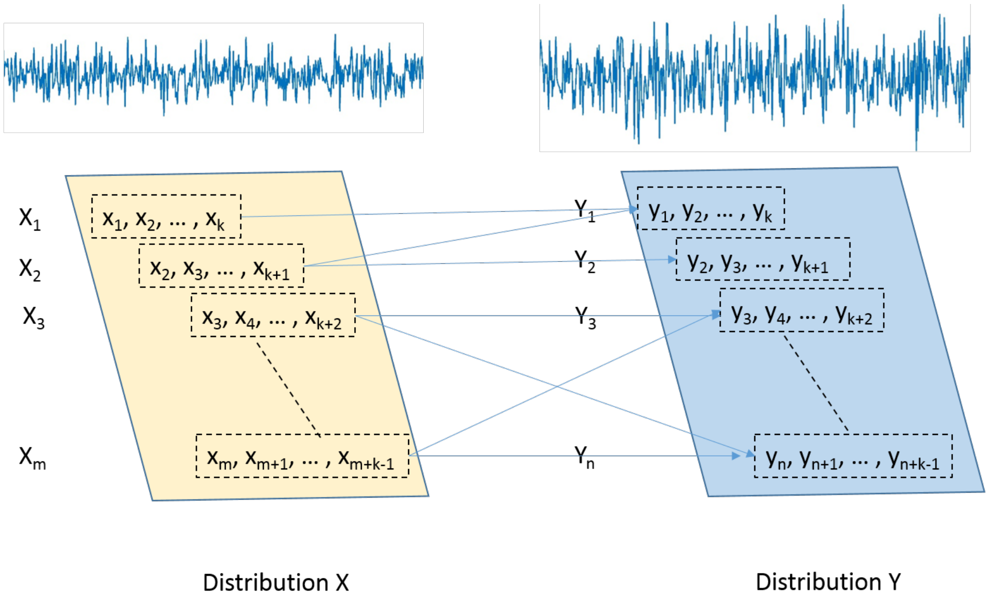

X_1
X_2
X_3
X_m
$x_1, x_2, \ldots, x_k$
$x_2, x_3, \ldots, x_{k+1}$
$x_3, x_4, \ldots, x_{k+2}$
$x_m, x_{m+1}, \ldots, x_{m+k-1}$
Y_1
Y_2
Y_3
Y_n
$y_1, y_2, \ldots, y_k$
$y_2, y_3, \ldots, y_{k+1}$
$y_3, y_4, \ldots, y_{k+2}$
$y_n, y_{n+1}, \ldots, y_{n+k-1}$
Distribution X
Distribution Y

Figure 5.3: Schematic representation of the data segmentation strategy

.

each element is a matrix, which can be rearranged as a vector, and then the vector norm can be used in a similar fashion.

5.3.2 Change detection within a single signal

When the change point is to be detected within a time series data without a reference data representing the "normal" condition, segmentation is performed within the data set itself. This is illustrated through Figure 5.4. To evaluate the change potential at a time instant, a same size window of m samples of data before and after this time instant is taken as two time series signals and the segmentation and distribution generation is performed following the procedure in Section 5.3.1. Finally, the KD score for that specific time instant is computed. This procedure is repeated for every time instant. Finally, a KD score plot can be generated and it will be further used as the basis for change detection.

The change detection capability of this approach is illustrated in Figure 5.5 on a signal with intermediate mean shift and a signal with intermediate variance change.

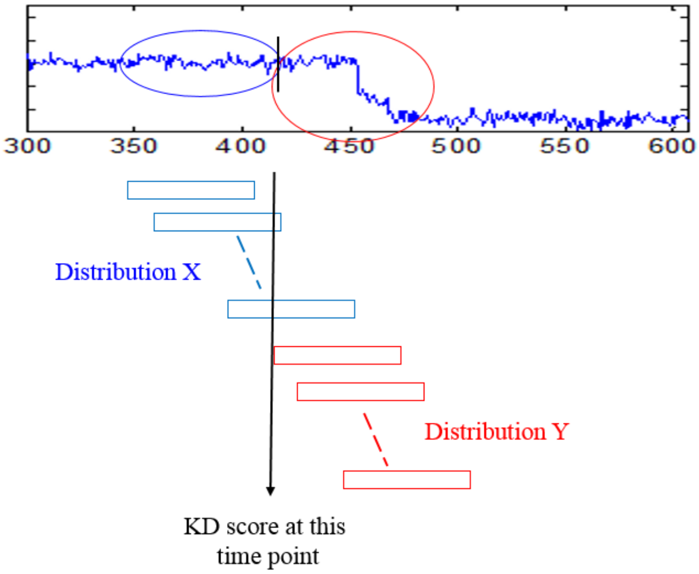
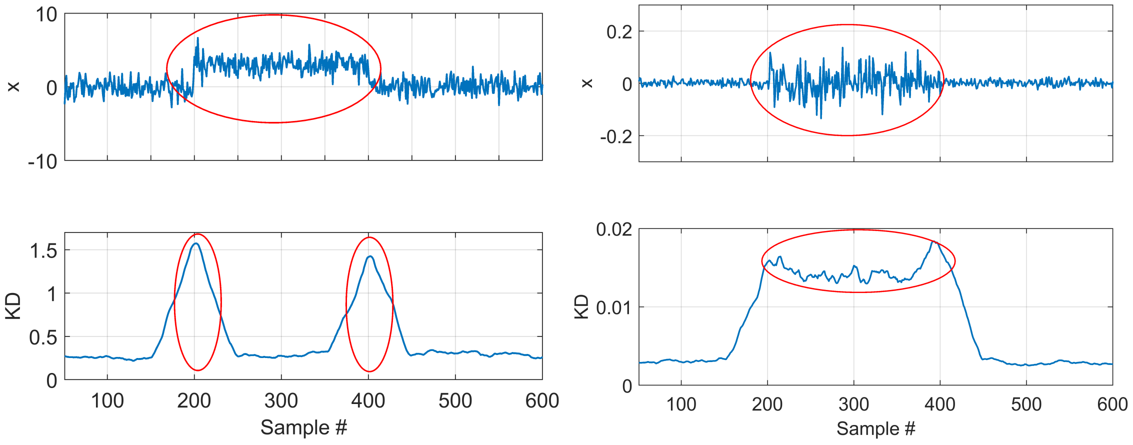


Figure 5.4: Data segmentation and distribution generation within single time series data

.

The mean shift detection capability of KD, calculated through the linear programming approach, is illustrated in Figure 5.5a. The signal x experiences a mean shift for a duration of 200 samples in the course of its time evolution. For each sample, the LP problem is solved for m samples of the signal preceding and succeeding the sample, including the current sample itself. It is observed that the points at which the signal mean shift occurs are clearly detected as significant peaks in the time evolution of KD. Next, the variance change detection capability of KD, calculated through the LP approach, is illustrated in Figure 5.5b. The signal x undergoes shows increased variance for a duration of 200 samples in the course of its time evolution. It is observed that the period of variance change of the signal is clearly detected in the time evolution of KD, as a sustained increase in its value.

The aforementioned strategy utilizes data from samples past and future with respect to the current data sample to calculate KD. In the case of on-line monitoring, the future samples are not available and so, detection is delayed. It is observed that



(a)

(b)

Figure 5.5: Change point detection performance of KD for data with (a) mean shift, and (b) variance change (right)

the larger the number of segments chosen, the longer it will take for the change to be detected. Hence, detection delay is found to be directly proportional to the number of segments s. Another parameter to be chosen in this approach is the number of data points k in each segment. Arifin *et al.* (2018) found that the KD score profile is smoother when the value of k is larger.

When the above strategy is used on a signal that undergoes a (faulty) magnitude variation in its course, only the point of change from normal process operation is detected by a significant rise in the KD value. However, once the past and future windows of m samples corresponding to the two distributions used for the KD calculation move into the fault-ridden signal region, the KD drops back to a negligible value, thereby providing a false positive that the signal has moved back into normal process operation. Therefore, to be able to use KD as a fault detection metric, a reference data set representing normal process operation must be used, against which the on-line testing data must be compared, to calculate the KD. In the next subsection, we investigate the change detection performance with a reference signal.

5.3.3 Change detection against a reference signal

In this section, the change detection using the KD metric is illustrated against a reference data set. The test signals are generated from five data distributions: Gaussian, Student's t, uniform, Gaussian mixture, and beta. We also evaluate the performance of the metric calculated through the linear programming approach with that calculated through the closed form expression.

For each distribution, two sets (D_1, D_2) of 1000 samples of pure noise were generated. For D_2, a constant bias type of change was added to each noise vector from sample number 300 to the end. D_1 serves as the reference data and D_2 serves as the test data. D_2 is illustrated in Figure 5.6. A moving window of $m = 50$ samples was employed on D_2 and the KD was computed between these samples (Y) and a chosen subset of the vectors from $D_1(X)$, by solving the LP problem at each sample point. Therefore, the test noise signal (with change) is compared against a reference signal, at each sample point. In addition, KD is also computed for the same subsets at each sample point using the closed-form expression. The change point detection performance through both approaches in contrasted in Figure 5.7.

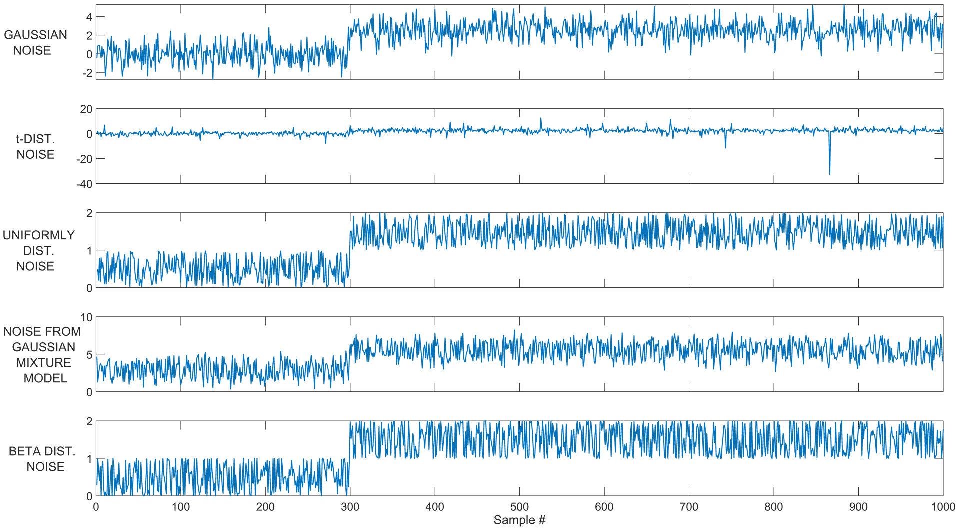


Figure 5.6: Time series data for pure noise signals from various distributions

Figures 5.7 (left panel) illustrate the change detection performance of KD computed using the LP approach, while Figures 5.7 (right panel) illustrate that of KD computed using the closed form expression. Both sets of results depict satisfactory change detection performance of the KD metric using a reference data set. While the first two distributions are elliptical distributions and the closed form KD is applicable, we also test the rest to see the performance on other type of distributions. The results indicate that the closed form expression of KD developed for data from multivariate Gaussian measures performs well for non-Gaussian data also. Therefore, for the proposed fault detection scheme in Sections 5.4 and 5.5, we utilize the closed form expression to compute KD.

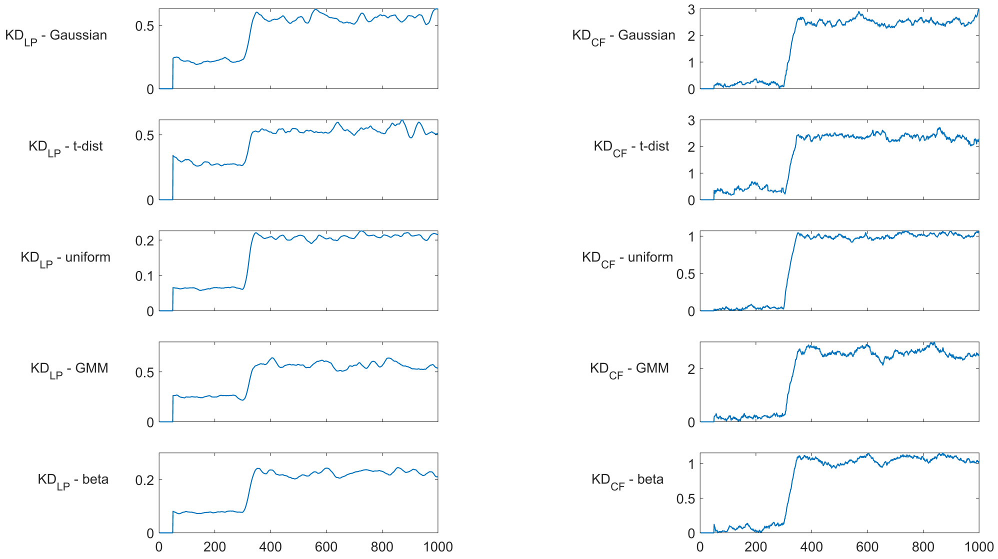


Figure 5.7: Time series evolution of KD for different distributions: (left panel) using LP approach, (right panel) using the closed form expression

In the subsequent section, a fault detection method using KD is presented. This method relies on the technique of online change point detection with reference signal obtained from normal process operations.

5.4 Fault detection using KD

In the realistic process monitoring problem, we often deal with multivariate process data where correlation exists. This correlation in the data can further be exposed by data-driven modelling such as Principal Component Analysis. Any changes that occur in the data signals are found to be amplified in the residual subspace and so, a fault detection framework that combines the proposed fault detection method and PCA is presented here.

5.4.1 Principal Component Analysis

PCA is a quantitative feature extraction method that provides insight into the underlying structure of a dataset (Venkatasubramanian *et al.* 2003b). PCA takes a multidimensional correlated data set and projects it into an uncorrelated space by maximizing the variance (of the original data) captured in each new dimension (Harrou *et al.* 2017b). As a result, a fewer number of the new dimensions are usually sufficient to capture the essence of the original dataset. PCA may be accomplished by the Singular Value Decomposition (SVD) of the covariance matrix (Σ) of the original dataset $X \in \mathbb{R}^{r \times c}$ as follows. Here, r refers to the number of $c-$dimensional data samples in X. It is to be noted that SVD is performed on the dataset that has been scaled to zero mean and unit variance.

$$X = USP^{\mathrm{T}} \tag{5.9}$$

The diagonal matrix S contains the eigenvalues of Σ in descending order, and the matrix P contains the orthonormal eigenvectors (column-wise), termed as loadings, associated with each eigenvalue in S. The transformed data is termed as the scores, and is obtained by the multiplying the data with the loadings. When the percentage of variance to be captured is specified as v, the PCA model can be obtained by retaining l principal components (PCs), where l is the number of PCs to be retained to capture v. Several techniques have been developed to compute the optimal number of PCs

(l) to be retained, such as the Scree plot, the cumulative percent variance (CPV) approach, the cross validation method, etc. In the proposed fault detection scheme, the CPV approach has been chosen to compute l. The CPV captured retaining p PCs at a time is calculated as,

$$CPV(p) = \frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{c} \lambda_i} \times 100 \tag{5.10}$$

The optimal number of PCs (l) is chosen as $l = p$ when,

$$CPV(p) \geq v \tag{5.11}$$

The model prediction for the chosen l-dimensional PCA model is obtained as follows.

$$\hat{X} = XP_l P_l^{\mathrm{T}} \tag{5.12}$$

Here, P_l refers to the l retained eigenvectors from the matrix P. The PCA-model residuals are calculated as,

$$E = X - \hat{X} \tag{5.13}$$

5.4.2 Conventional PCA-based fault detection scheme

The conventional PCA-based fault detection scheme utilizes T^2 and Q statistics to analyse the principal and residual subspaces respectively.

The Squared Prediction Error (SPE) or the Q statistic is a quantification of the residual subspace of the PCA model of a dataset. It is a metric that accounts for the amount of variance that is not captured by the chosen l-dimensional PCA model. The Q statistic for each c-dimensional sample of data is computed as follows.

$$Q = e^{\mathrm{T}} e \tag{5.14}$$

Here, e denotes the PCA model residual for the chosen sample of data. The threshold for the Q statistic, denoted by Q_α, is computed using the formula specified by Jackson and Mudholkar (1979). Here, c_α is the standard normal variate with confidence level $(1 - \alpha)$.

$$Q_\alpha = \theta_1 \left(\frac{h_0 c_\alpha \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right)^{\frac{1}{h_0}} \tag{5.15}$$

where $\theta_i = \sum_{j=l+1}^{c} \lambda_j^i$ and $h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$.

The Hotelling's T^2 statistic is a quantification of the principal subspace of the PCA model of a dataset. This metric accounts for the variance that is captured by the chosen l-dimensional PCA model. The T^2 statistic for each c-dimensional sample of data is computed as follows (Villegas *et al.* 2010).

$$T^2 = X^{\mathrm{T}} P_l S_l^{-2} X P_l^{\mathrm{T}} \tag{5.16}$$

Here, S_l is a diagonal matrix that contains the eigenvalues associated with the l retained PCs. The threshold for the T^2 statistic, denoted by T_α^2, is computed as follows (Villegas *et al.* 2010).

$$T_\alpha^2 = \frac{(n-1)l}{n-l} F_{l,n-l,\alpha} \tag{5.17}$$

The $l-$dimensional PCA model is developed for scaled training data and the thresholds Q_α and T_α^2 are calculated. The model is used to obtain data estimates, and Q and T^2 statistics for the testing data set. The violation of the Q and/or T^2 statistics is interpreted as the detection of an abnormality in the dataset. The flow diagram of this conventional scheme is illustrated in Figure 5.8.

5.4.3 Proposed PCA-based fault detection scheme with KD metric

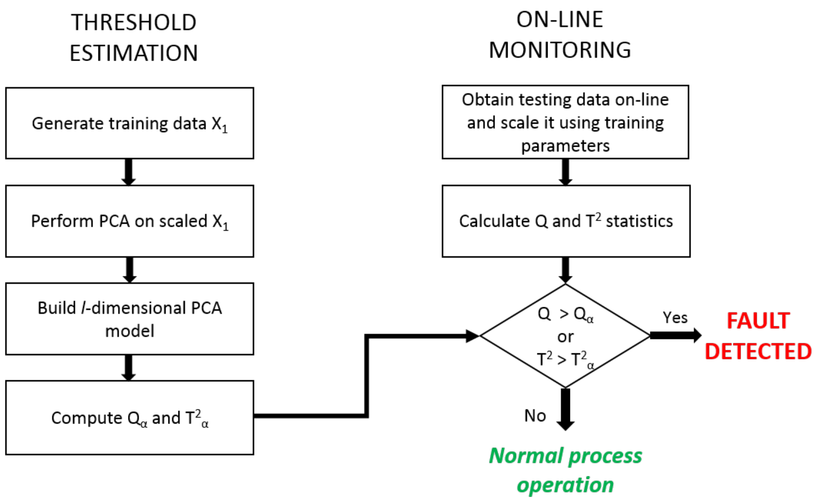

THRESHOLD ESTIMATION
Generate training data X_1
Perform PCA on scaled X_1
Build *l*-dimensional PCA model
Compute Q_α and T^2_α
ON-LINE MONITORING
Obtain testing data on-line and scale it using training parameters
Calculate Q and T^2 statistics
$Q > Q_\alpha$ or $T^2 > T^2_\alpha$
Yes
FAULT DETECTED
No
Normal process operation

Figure 5.8: Flow diagram representation of the conventional PCA-based fault detection scheme

The PCA based fault detection scheme using the KD metric follows the same basic outline as that of the conventional scheme, with a few modifications. The algorithm for the PCA-based fault detection scheme using KD is outlined below, and the corresponding flow diagram is illustrated in Figure 5.9.

Threshold estimation:

1. Obtain 2 sets of normal process data – D_1 and D_2. Scale the data to zero mean and unit variance

2. Build separate l–dimensional PCA models for both datasets, for v % variance to be captured

3. Obtain PCA model residuals for both datasets – R_1 and R_2

4. Set R_1 as x. To compute KD for each sample, employ a moving window of m samples on R_2 and set this as y

5. Compute KD between x and y for each sample using the closed-form expression

6. Compute the mean (μ_{KD}) and standard deviation (σ_{KD}) of KD for normal operation and set the threshold as $h = \mu_{\text{KD}} + 3\sigma_{\text{KD}}$

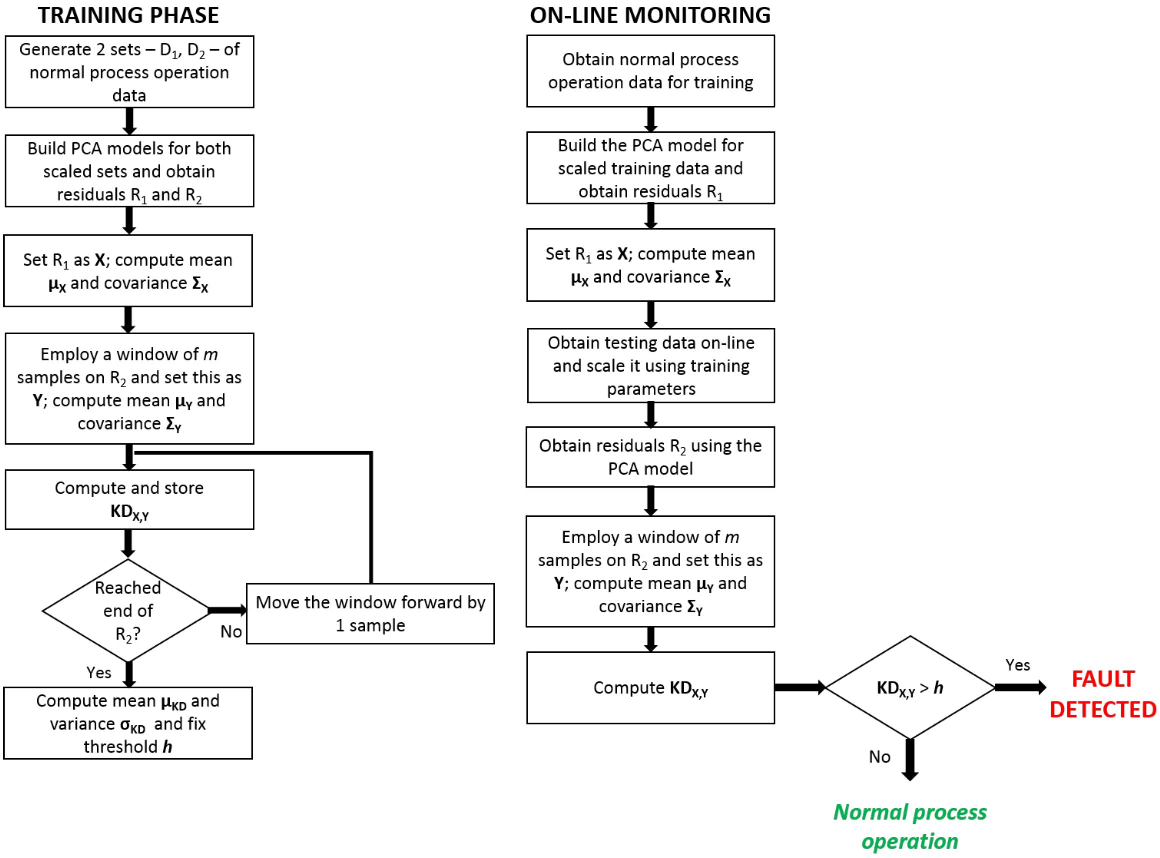

TRAINING PHASE
Generate 2 sets – D1, D2 – of normal process operation data
Build PCA models for both scaled sets and obtain residuals R1 and R2
Set R1 as X; compute mean μX and covariance ΣX
Employ a window of m samples on R2 and set this as Y; compute mean μY and covariance ΣY
Compute and store KDX,Y
Reached end of R2?
No
Move the window forward by 1 sample
Yes
Compute mean μKD and variance σKD and fix threshold h
ON-LINE MONITORING
Obtain normal process operation data for training
Build the PCA model for scaled training data and obtain residuals R1
Set R1 as X; compute mean μX and covariance ΣX
Obtain testing data on-line and scale it using training parameters
Obtain residuals R2 using the PCA model
Employ a window of m samples on R2 and set this as Y; compute mean μY and covariance ΣY
Compute KDX,Y
KDX,Y > h
Yes
FAULT DETECTED
No
Normal process operation

Figure 5.9: Flow diagram representation of the proposed PCA-based fault detection scheme with the KD metric

On-line monitoring:

1. Obtain a set of normal process data D_{main}. Scale the data to zero mean and unit variance

2. Build an $l-$dimensional PCA model for D_{main}, for v % variance to be captured

3. Obtain PCA model residuals for $D_{\text{main}} - R_{\text{main}}$.

4. Obtain testing dataset D_{test}. Scale the data using the normal process parameters from step 1.

5. Obtain PCA model residuals for $D_{\text{test}} - R_{\text{test}}$.

6. Set R_{main} as x. To compute KD for each sample, employ a moving window of m samples on R_{test} and set this as y.

7. Compute KD between x and y for each sample using the closed-form expression.

8. If KD > h, fault is detected.

Notice that the normal process data in the online monitoring stage (D_{main}) can be the any of the two sets used in training phase. It can be also another independent new set of normal data. In the proposed algorithm, the parameters v and m are given as inputs by the operator. The choice of v and m is system-specific, and is found to be largely dependent on the nature and noise content of the data. Parameter v is chosen to obtain an adequate PCA model of the system, leaving the noisy characteristics of the data out of the model. This enables us to obtain significant residuals even in the presence of smaller magnitude faults, and thus, enables detection of more sensitive faults. The size of the normal data sets used in the algorithm and the choice of moving window size m, are made to obtain an adequate estimate of the mean and covariance information of the signal. As a result, m is dependent on the type and amount of noise in the signal: if the signal is significantly noise-ridden, a higher choice of m might be required to obtain a better estimate of the mean and variance.

The proposed method is based on the calculation of KD between two sets of residuals. The KD is a metric to quantify the distance between probability distributions. In this sense, it is more appropriate to classify it as a distance metric than a correlation metric.

5.5 Case study

5.5.1 Case 1: Synthetic data

To illustrate the fault detection capability of the KD metric, the synthetic dataset from Chapter 2 of the thesis of Harmouche (2014) was generated. The dataset contains 1000 samples of 8 variables, out of which 3 are independent. In this section, we evaluate the fault detection capability of the KD metric for data corrupted with (i) Gaussian noise, and (ii) $t-$distributed noise. A signal to noise ratio (SNR) of 35 was maintained for both cases, as in the original thesis. The equations for generation of the raw dataset are listed in 5.18. Here, n refers to the sample number between 1 and 1000, and N refers to the total number of samples (in this case, 1000).

$$x_1 = 1 + \sin(0.1n) \tag{5.18a}$$

$$x_2 = 2\cos^3\left(\frac{n}{4}\right)\exp\left(\frac{-n}{N}\right) \tag{5.18b}$$

$$x_3 = \log_{10}(\chi_2^2) \tag{5.18c}$$

$$x_4 = x_1 + x_2 \tag{5.18d}$$

$$x_5 = x_1 - x_2 \tag{5.18e}$$

$$x_6 = 2x_1 + x_2 \tag{5.18f}$$

$$x_7 = x_1 + x_3 \tag{5.18g}$$

$$x_8 \sim \mathbb{N}(0,1) \tag{5.18h}$$

5.5.1.1 Data corrupted with Gaussian noise

The proposed PCA model residual-based fault detection scheme utilizing the KD metric was applied on data corrupted with Gaussian noise, shown in Figure 5.10. The detection ability of the metric was investigated for 3 types of faults – a sustained bias, an incipient fault, and intermittent faults. 4 PCs were retained for $v = 95\%$ of the original variance captured. A sustained bias of 0.25 in x_1, a drift with slope 0.001

in x_1 and intermittent faults of magnitude 0.2, 0.5 and 0.8 on variables x_1, x_4, and x_6 were applied in each case, respectively. All faults were introduced at sample point 300. A moving window of m = 50 samples was used. The fault detection results for KD, as well as Q and T^2 statistics are illustrated in Figures 5.11 - 5.13.

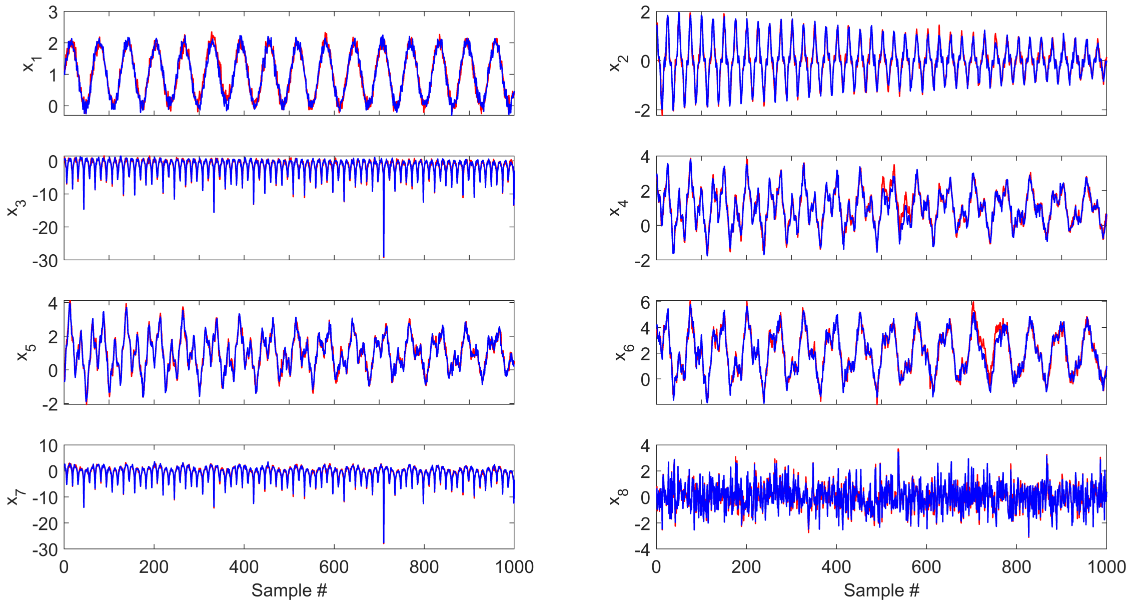


Figure 5.10: Synthetic reference data (blue) vs testing data with intermittent faults (red)

From the results, the following observations can be made. Firstly, it is evident that the fault detection capability of the KD metric is superior to that of the Q and T^2 statistics. KD gives no false alarms prior to the introduction of the fault and exhibits sustained violation of the threshold after the fault is introduced. In the case of the Q statistic, the noisy nature of the data leads to false alarms and missed detection. Furthermore, it is observed that the T^2 statistic does not respond to the small magnitude of faults introduced, whereas the response of the KD metric is appreciable. Therefore, it is inferred that the KD metric offers better fault detection capability when compared to the conventional metrics, even when data is significantly corrupted with Gaussian noise. It is to be noted that the KD statistic detects the

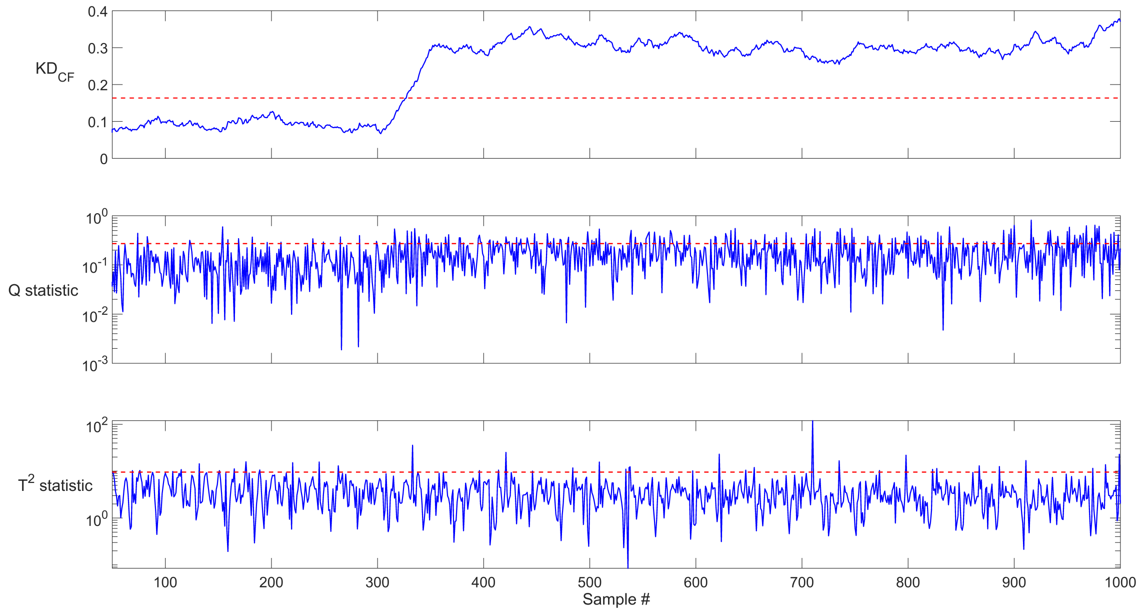


Figure 5.11: Detection results for data with a sustained bias of +0.25 in x_1 (corrupted with Gaussian noise)

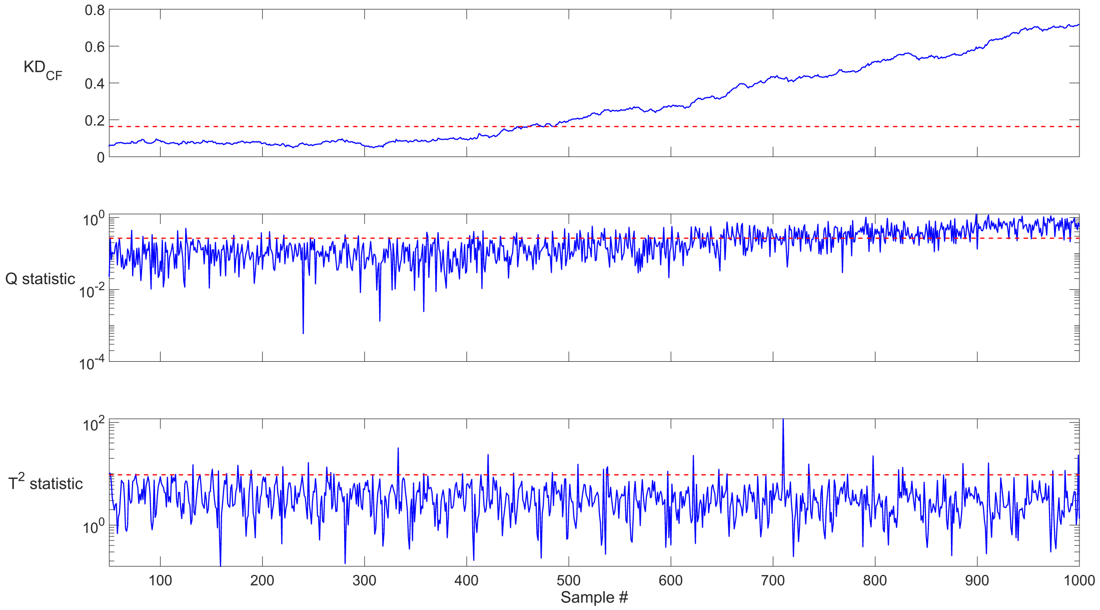


Figure 5.12: Detection results for data with an incipient anomaly of slope 0.001 in x_1 (corrupted with Gaussian noise)

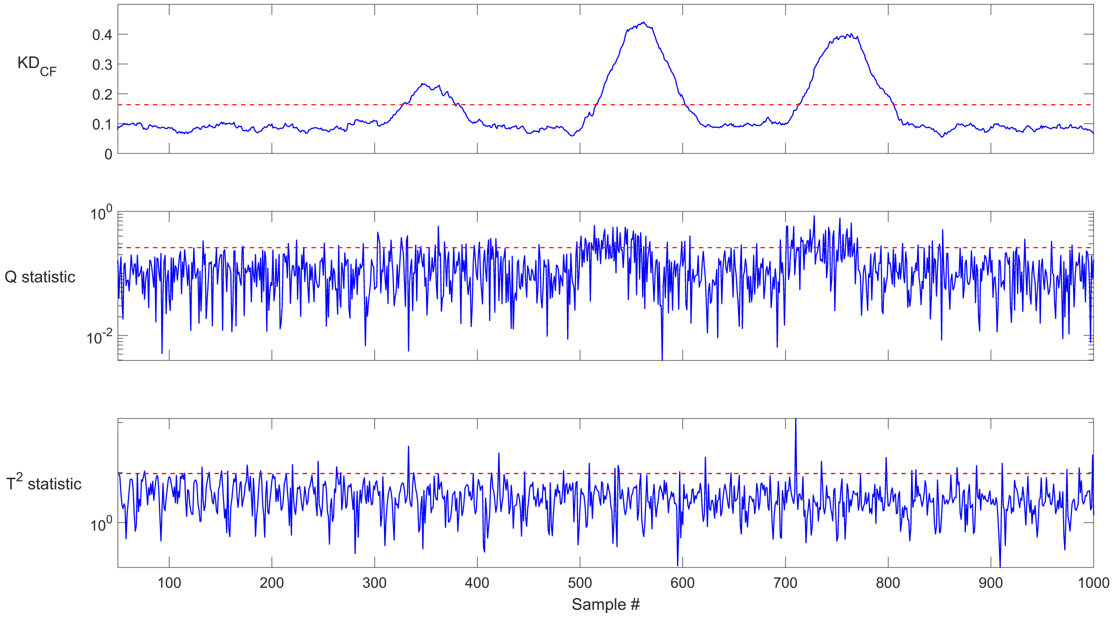


Figure 5.13: Detection results for data with multiple intermittent faults of +0.2, +0.5 and +0.8 in x_1,x_4 and x_6 (corrupted with Gaussian noise)

drift fault with a delay. This is attributed to the small slope of the drift (0.001).

5.5.1.2 Data corrupted with $t-$distributed noise

The proposed PCA model residual-based fault detection scheme utilizing the KD metric was applied on data corrupted with $t-$distributed noise with 5 degrees of freedom. The purpose of using $t-$distribution is to test the detection performance while the data contains a few outliers, for which a $t-$distribution is more appropriate. The detection ability of the metric is investigated for the same types and magnitude of faults as in Section 5.5.1.1. In this case also, 4 PCs were retained for $v = 95\%$ of the original variance captured. A moving window of m = 50 samples was used. The fault detection results for KD, as well as Q and T^2 statistics are illustrated in Figures 5.14 - 5.16.

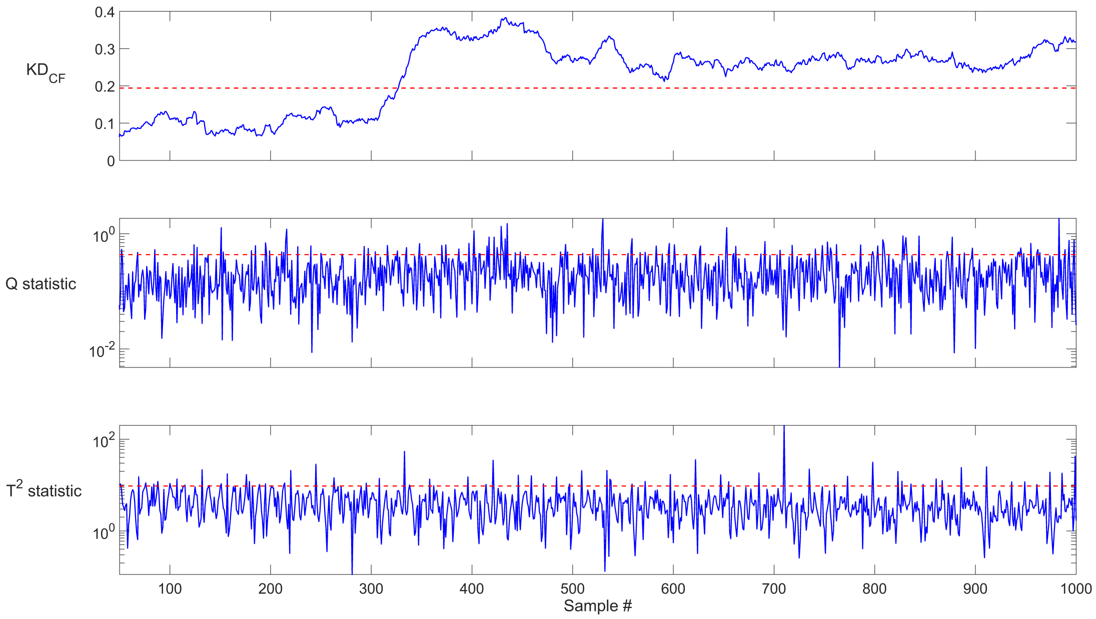


Figure 5.14: Detection results for data with a sustained bias of +0.25 in x_1 (corrupted with t−distributed noise)

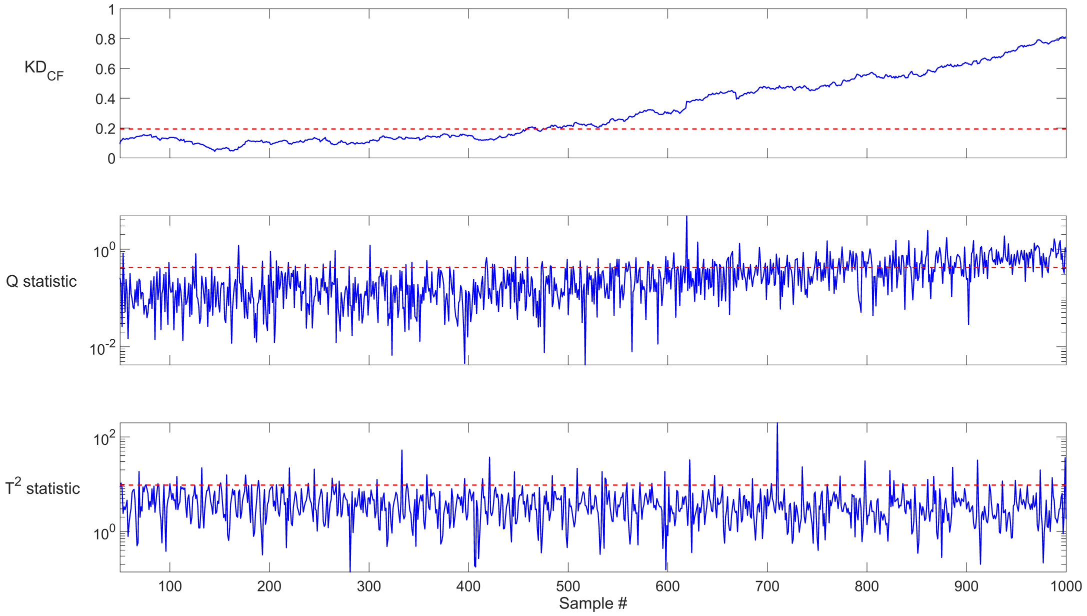


Figure 5.15: Detection results for data with an incipient anomaly of slope 0.001 in x_1 (corrupted with t−distributed noise)

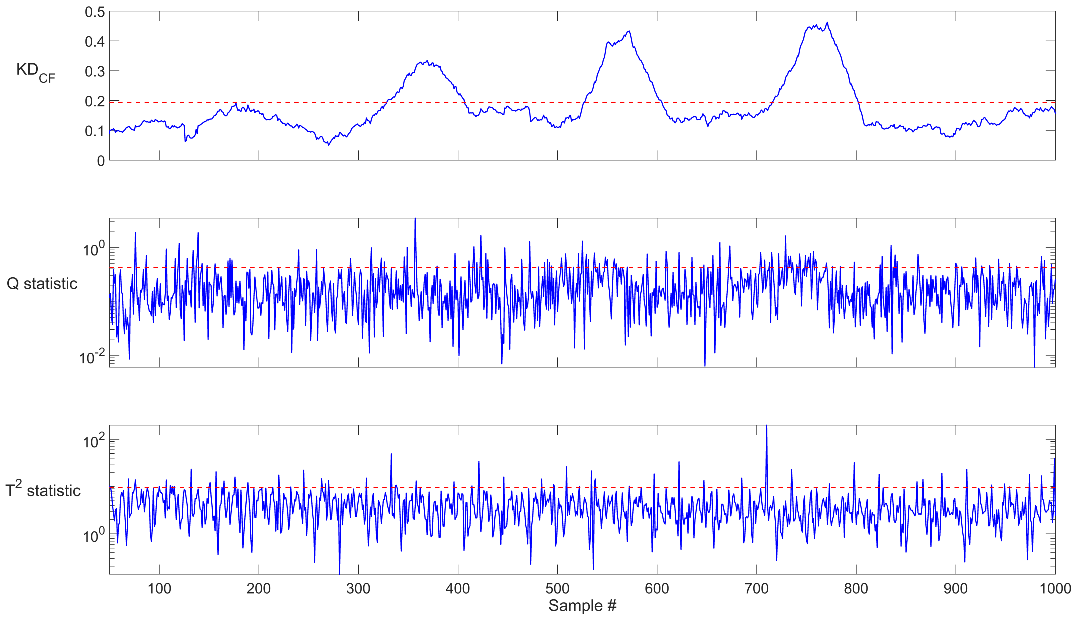


Figure 5.16: Detection results for data with multiple intermittent faults of +0.2, +0.5 and +0.8 in x_1,x_4 and x_6 (corrupted with $t-$distributed noise)

From the results, it is observed that the fault detection performance of the KD metric remains largely unchanged, when the nature of the noise in the data set is changed from Gaussian to non-Gaussian. $t-$distributed noise in the data is characterized mainly by random large spikes in the measurements. It is observed that the KD metric still offers a superior performance, when compared with the Q and T^2 statistics, with no false alarms and no missed detection. The performance of the Q and T^2 statistics still remains unsatisfactory, being largely affected by the noise content in the signals, giving rise to numerous false alarms and missed detections. It is to be noted that the KD statistic detects the drift fault with a delay of approximately 170 samples.

5.5.2 Case 2: Continuous Stirred Tank Heater (CSTH) simulation

The data for this case study is generated from a simulation developed for the Continuous Stirred Tank Heater (CSTH) setup (Thornhill *et al.* 2008). This system comprises two stirred tank heater units to which cold water is supplied (u_1 and u_2). Inputs u_4 and u_5 feed the heater inputs Q_1 and Q_2 to the tanks, and input u_3 recycles water from tank 2 to 1. The output variables of interest are the tank outlet temperatures T_1 and T_2 measured in K, and the liquid level in tank 2 measured in m. Flow rates are provided in $\mathrm{m}^3\mathrm{s}^{-1}$ and heater inputs in Js^{-1}. The system is described by equations 5.19, and a schematic representation is provided in Figure 5.17. The nominal model parameters and steady state values are given in Table 5.1.

$$V_1 \frac{dT_1}{dt} = F_1(u_1)(T_c - T_1) + F_R(u_3)(T_2 - T_1) + \frac{Q_1(u_4)}{\rho C_p} \tag{5.19a}$$

$$A_2 h_2 \frac{dT_2}{dt} = \begin{array}{c} F_1(u_1)(T_1 - T_2) + F_2(u_2)(T_c - T_2) - F_R(u_3)(T_2 - T_1) \\ + \frac{1}{\rho C_p}\left[Q_2(u_5) - 2\pi r_2 h_2 U(T_2 - T_a)\right] \end{array} \tag{5.19b}$$

$$A_2 \frac{dh_2}{dt} = F_1(u_1) + F_2(u_2) - F_{\text{out}}(h_2) \tag{5.19c}$$

$$F_{\text{out}}(h_2) = (0.1 \times 10^{-3})\sqrt{0.406h_2^3 + 0.8061h_2^2 - 0.01798h_2 + 0.1054} \tag{5.19d}$$

$$F_1(u_1) = (42379u_1 - 456.85u_1^2 + 8.0368u_1^3) \times 10^{-11} \tag{5.19e}$$

$$F_2(u_2) = (196620u_2 - 8796.8u_2^2 + 190.64u_2^3 - 1.294u_2^4) \times 10^{-11} \tag{5.19f}$$

$$F_R(u_3) = 2u_3\left(\frac{1}{3600}\right) \times 10^{-3} \tag{5.19g}$$

$$Q_1(u_4) = 7.9798u_4 + 0.9893u_4^2 - (7.3 \times 10^{-3}u_4^3) \tag{5.19h}$$

$$Q_2(u_5) = 104 + 14.44u_5 + 0.96u_5^2 - (8 \times 10^{-3}u_5^3) \tag{5.19i}$$

The inputs u_2, u_4, and u_5 are supplied to the system as pseudo-random binary signals, while u_1 and u_3 are steady state signals corrupted with zero mean Gaussian noise e passed through the first-order filter depicted below (set $\alpha = 0.95$). The simulation was run for 1000 samples of data, collected for normal process operation.

$$\frac{u_k(n)}{e_k(n)} = \frac{1 - \alpha}{1 - \alpha q^{-1}}, \quad \text{where} \quad k = 1, 3 \tag{5.20}$$

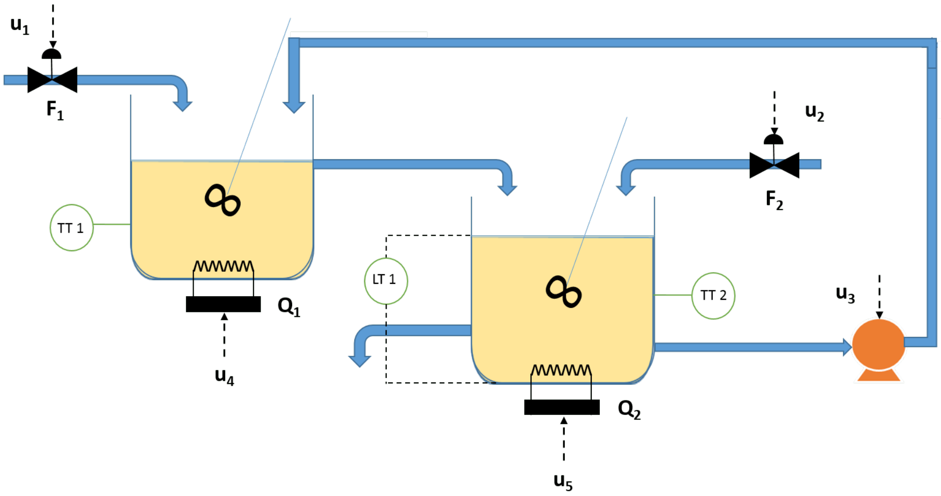


Figure 5.17: Schematic representation of the CSTH setup

Table 5.1: Model parameters and steady state operating conditions for the CSTH simulation

Parameter	Description	Value
V_1	Volume of tank 1	1.75×10^{-3} m^3
A_2	Cross sectional area of tank 2	7.854×10^{-3} m^2
r_2	Radius of tank 2	0.05 m
U	Heat transfer coefficient	235.1 W/m^{2}K
T_c	Cooling water temperature	30 ^{0}C
T_a	Ambient temperature	25 ^{0}C
u_1	Flow F_1 (% input)	50%
u_2	Flow F_2 (% input)	50 %
u_3	Flow F_R (% input)	50 %
u_4	Heat input Q_1 (% input)	60 %
u_5	Heat input Q_2 (% input)	50 %

5.5.2.1 Data corrupted with Gaussian noise

The proposed PCA model residual-based fault detection scheme using the KD metric was applied on CSTH data corrupted with Gaussian noise. It is to be noted that noise (SNR = 35) is added only to the output variables – $T_1, T_2,$ and h_2 – in order to simulate real sensor measurements. The detection ability of the metric was investigated for 3 types of faults – a sustained bias, an incipient fault, and intermittent faults. 6 PCs were retained for $v = 95\%$ of the original variance captured. A sustained bias of +1 ^{0}C in T_1, a drift with slope 0.002 in T_2 and intermittent faults of magnitude +1.25 ^{0}C in T_1, +2 ^{0}C in T_2, and 0.02 m in h_2, were applied in each case, respectively. All faults were introduced at sample point 300. A moving window of m = 50 samples was used. The fault detection results for KD, as well as Q and T^2 statistics are illustrated in Figures 5.18 - 5.20.

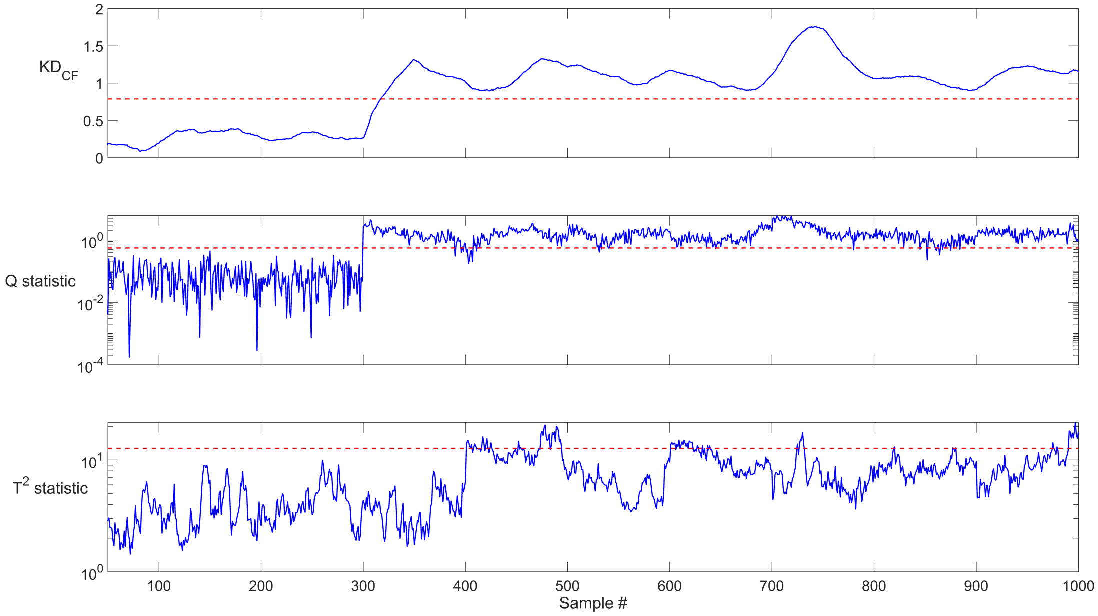


Figure 5.18: Detection results for data with a sustained bias of +1.25 ^{0}C in T_1 (corrupted with Gaussian noise)

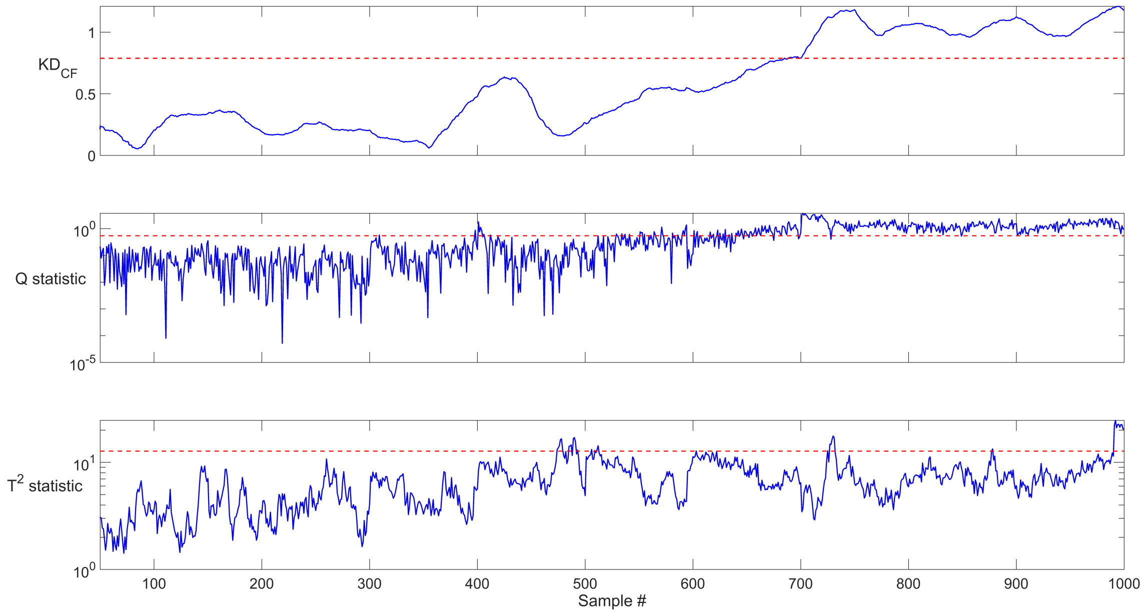


Figure 5.19: Detection results for data with an incipient anomaly of slope 0.002 in T_2 (corrupted with Gaussian noise)

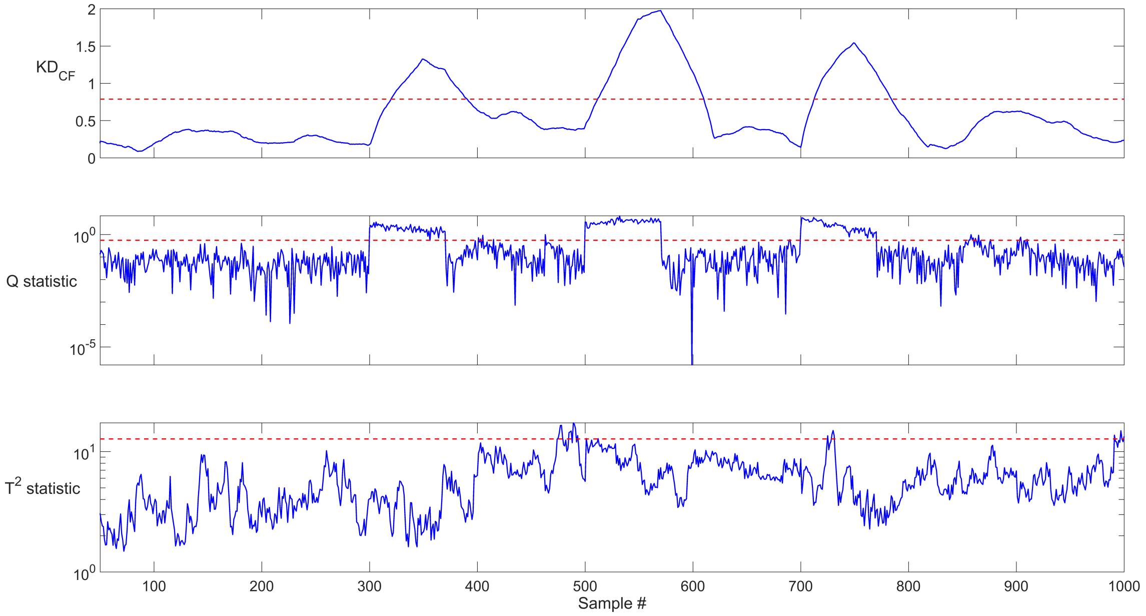


Figure 5.20: Detection results for data with multiple intermittent faults of +1.25 ^{0}C, +2 ^{0}C and +0.02 m in T_1, T_2 and h_2 (corrupted with Gaussian noise)

From the results, it is evident that the fault detection performance of the KD metric on PCA model-based residuals is far superior to that of the T^2 statistic, and also the Q statistic. The time evolution of the KD metric gives a smooth detection profile, with no missed detection during the fault period, and no false alarms before the fault is initiated. In contrast, the Q statistic appears to be heavily susceptible to missed detection, and the T^2 statistic shows low sensitivity to fault magnitude. It is to be noted that the KD statistic detects the drift fault with a delay of approximately 400 samples.

5.5.2.2 Data corrupted with t–distributed noise

The proposed PCA model residual-based fault detection scheme utilizing the KD metric was applied on CSTH data corrupted with t–distributed noise with 5 degrees of freedom, as illustrated in Figure 5.21. The detection ability of the metric is investigated for the same types and magnitude of faults as in Section 5.5.2.1. In this case also, 6 PCs were retained for $v = 95\%$ of the original variance captured. A moving window of m = 50 samples was used. The fault detection results for KD, as well as Q and T^2 statistics are illustrated in Figures 5.22 - 5.24. From the results, it is observed that the KD metric does indeed perform better than the Q and T^2 statistics for fault detection, irrespective of the nature of noise present in the signal. It is to be noted that the KD statistic detects the drift fault with a delay of approximately 400 samples.

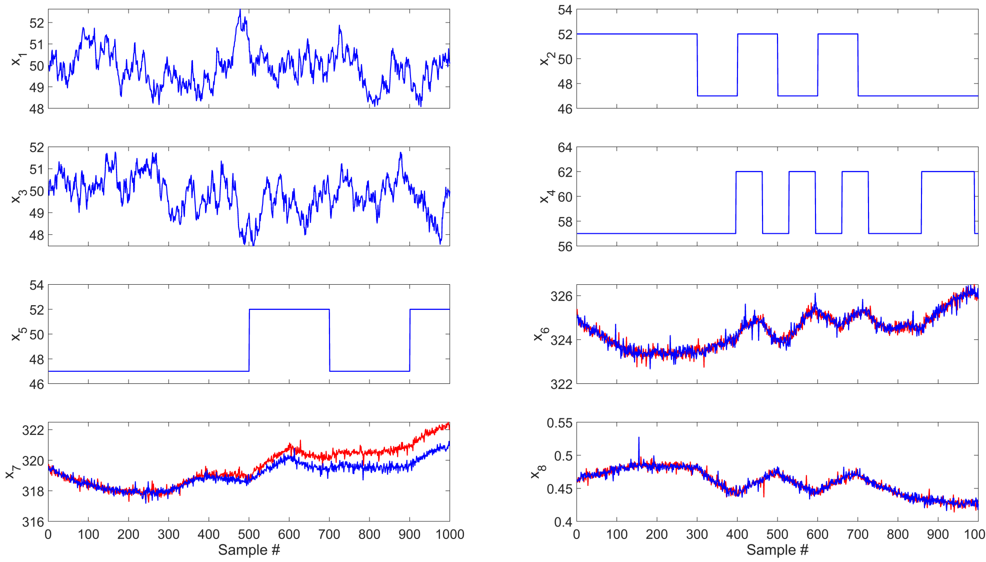


Figure 5.21: CSTH reference data (blue) vs testing data with incipient anomaly (red)

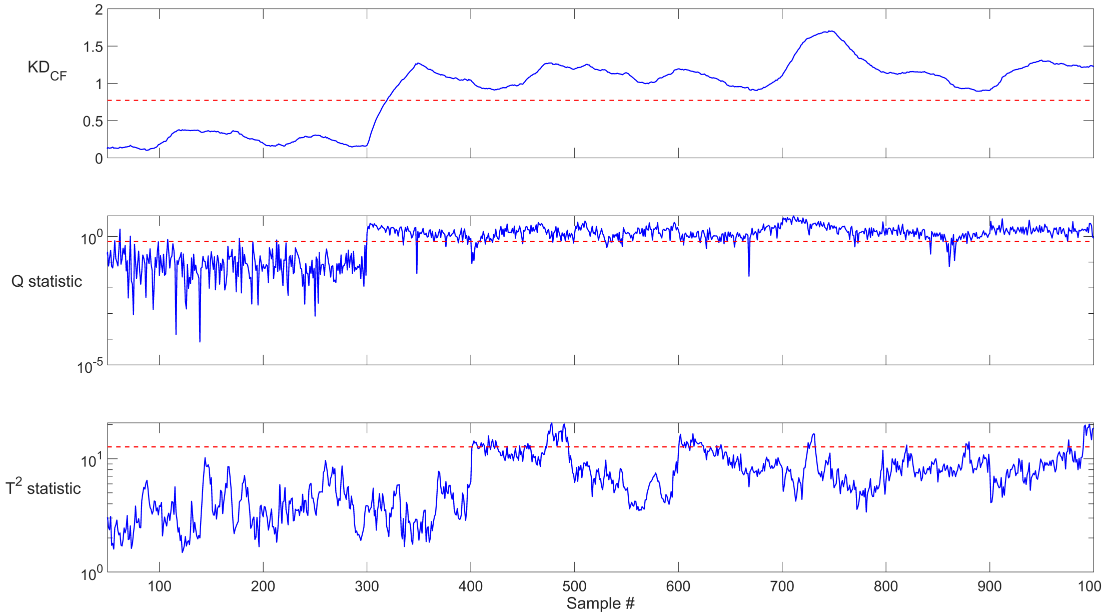


Figure 5.22: Detection results for data with a sustained bias of +1.25 ^{0}C in T_1 (corrupted with t−distributed noise)

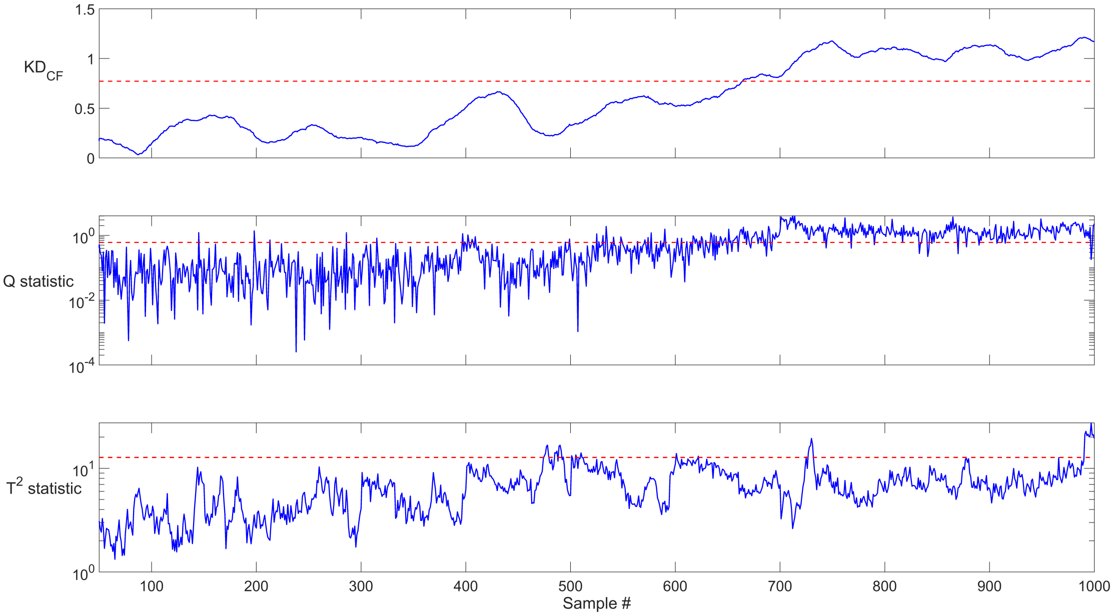


Figure 5.23: Detection results for data with an incipient anomaly of slope 0.002 in T_2 (corrupted with t–distributed noise)

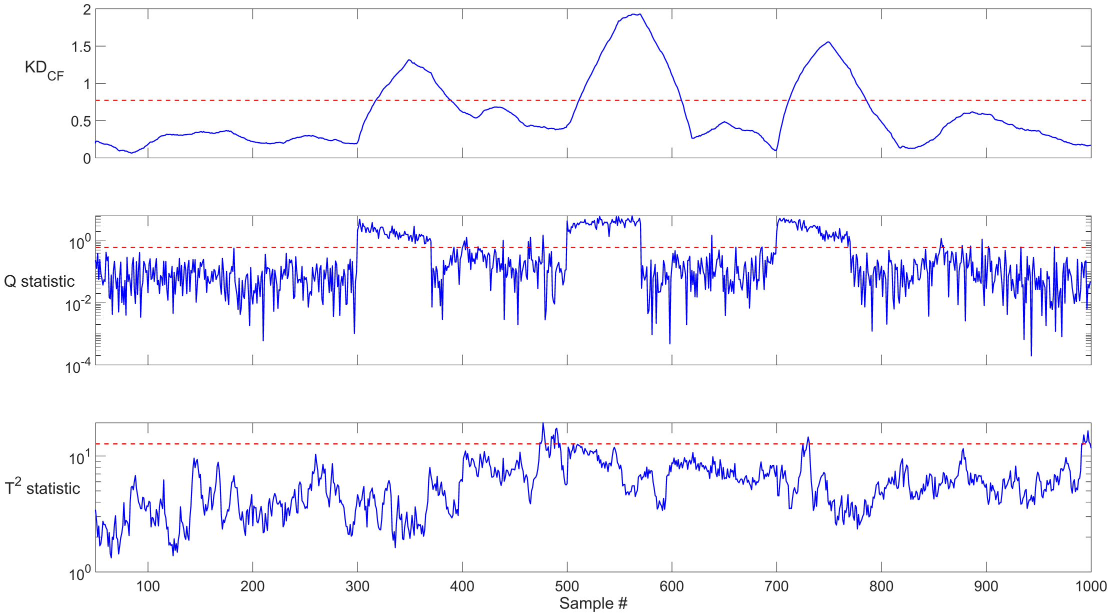


Figure 5.24: Detection results for data with multiple intermittent faults of +1.25 ^{0}C, +2 ^{0}C and +0.02 m in T_1, T_2 and h_2 (corrupted with t–distributed noise)

5.5.3 Tennessee Eastman process

The performance of the proposed fault detection algorithm using the KD metric with PCA model residuals is illustrated on the Tennessee Eastman process (TEP) simulation, a benchmark setup in the process control and monitoring research community. The TEP simulation is modelled on an actual process under the Eastman Chemical Company at Kingsport, Tennessee, U.S.A and the problem was formulated originally by Downs and Vogel (1993). In this process, four gaseous reactants are utilized to produce two products and one by-product through four irreversible and exothermic (approximately) first-order reactions with respect to reactant concentration, which include the presence of an inert compound. In total, the process involves eight chemical compounds and all products are obtained in the liquid phase.

The process comprises five major units – a reactor, a product condenser, a vapour-liquid separator, a recycle compressor and a product stripper. A complete description of the process can be found at Downs and Vogel (1993). The process comprises 41 measured variables (XMEAS), 22 of which are sampled continuously and the rest sampled at different intervals, and 12 manipulated variables (XMV). The bases case values, nominal operating conditions and steady state information is recorded in Downs and Vogel (1993).

The fault detection performance of the proposed algorithm in Section 5.4 is illustrated on four examples of faults: Fault 1 (step in A/C feed ratio in Stream 4), Fault 2 (random variation in condenser cooling water inlet temperature), Fault 3 (slow drift in reaction kinetics) and Fault 4 (sticking reactor cooling water valve). For all four cases, 36 PCs were retained from the 52, for $v = 95\%$ of the variance captured. A moving window of m = 50 samples was used. The fault detection results using KD are contrasted with those of the Q and T^2 statistics, and are illustrated in Figures

5.25 - 5.28. From the results, the following observations can be made. In all fault cases explored, it is noted that the KD metric offers a smoother time evolution. Additionally, the evolution of the KD metric does not seem to be dependent on the extremely fault-ridden profile of the faulty variable itself, as in the case of the Q and T^2 statistics in Figure 5.28. The KD metric, while containing some detection delay, offers a profile with no false alarms prior to the fault, and no false positives after the fault is introduced. In this context, the detection delay of the KD metric is found to be approximately 20 samples for Faults 1 and 2, 60 samples for Fault 3, and 15 samples for Fault 4. Furthermore, from the results, it is clear that the KD metric offers a large detection magnitude, that is, the KD when fault is detected is well above the threshold; comparing the KD profiles with those of the Q and T^2 statistics, it is observed that the latter statistics do not possess a large detection magnitude in this case study. Therefore, we infer that the KD metric would be more sensitive than the Q and T^2 statistics, and offer better detection for smaller magnitude faults also.

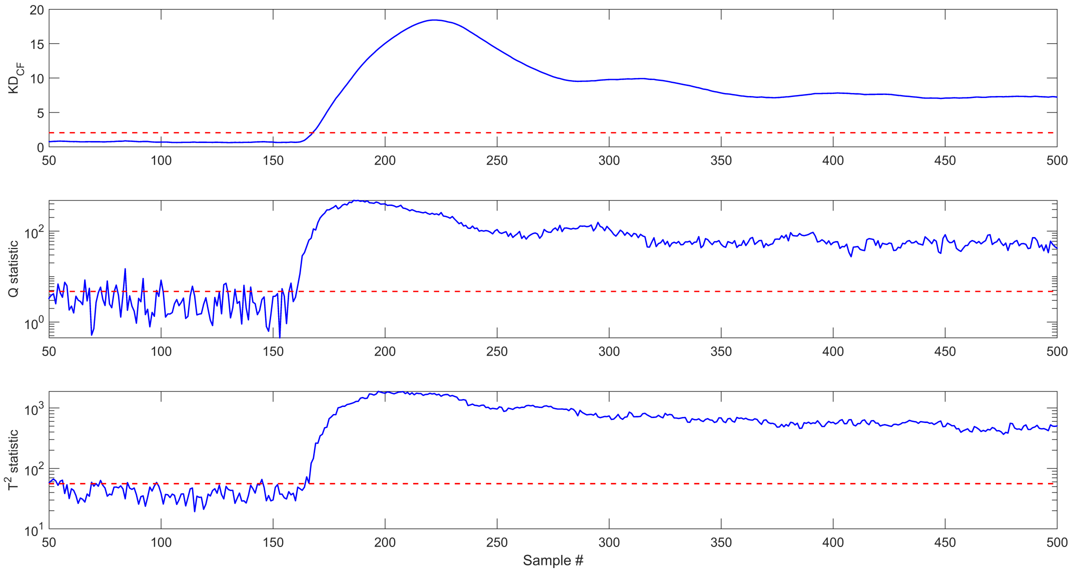


Figure 5.25: TEP with step in A/C feed ratio in Stream 4 (Fault 1)

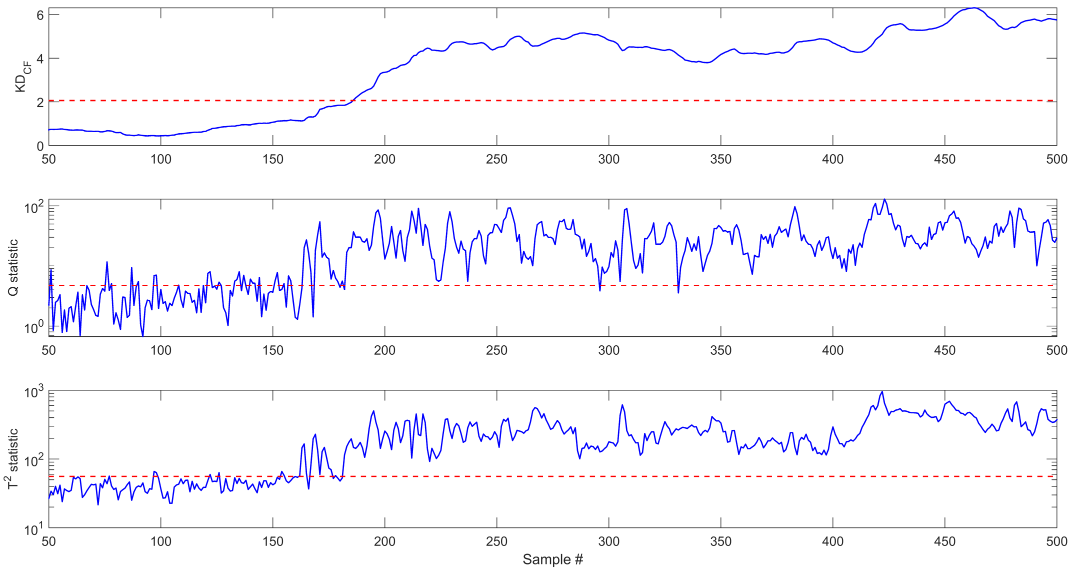


Figure 5.26: TEP with random variation in condenser cooling water inlet temperature (Fault 2)

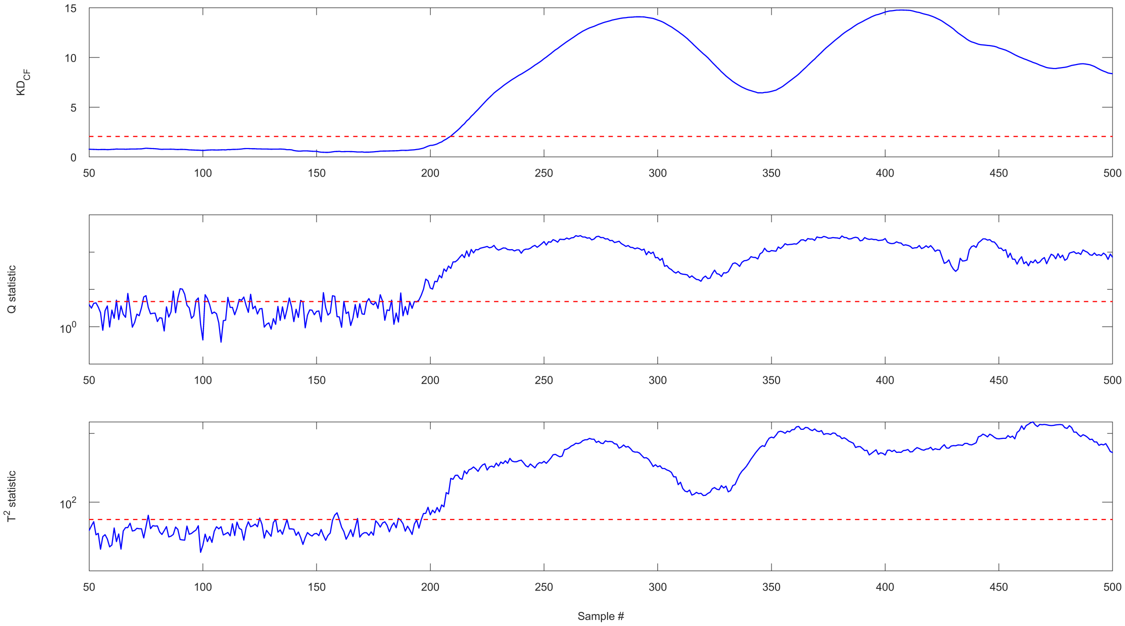


Figure 5.27: TEP with slow drift in reaction kinetics (Fault 3)

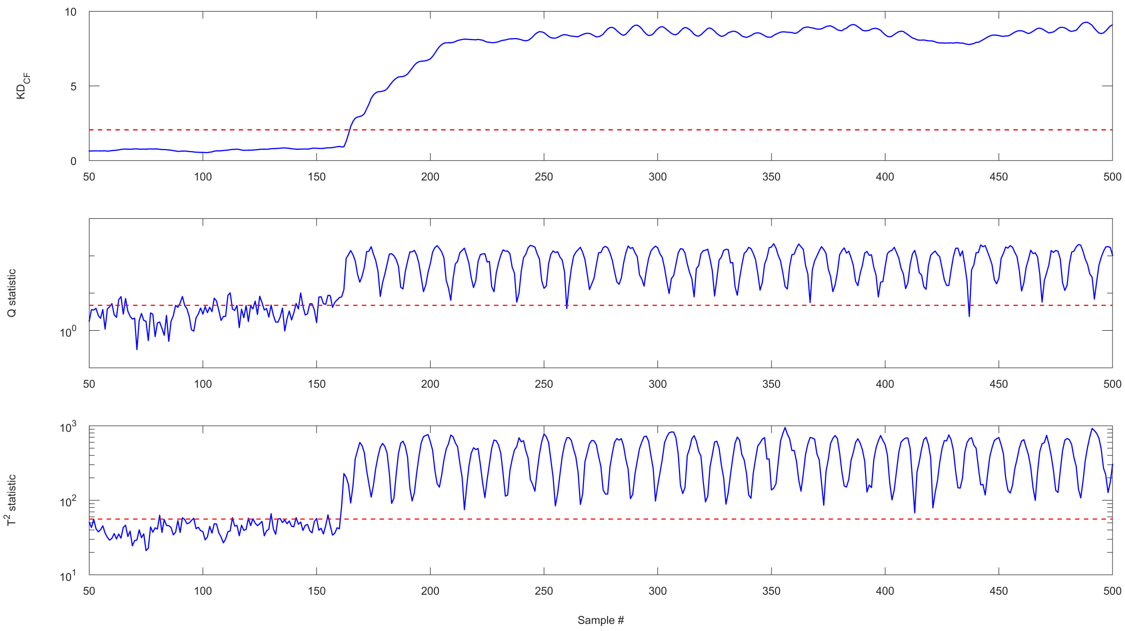


Figure 5.28: TEP with sticking reactor cooling water valve (Fault 4)

5.6 Conclusions

In this chapter, a novel change and fault detection scheme using the Kantorovich Distance metric has been proposed. The change detection capability of the metric has been evaluated for data sampled from different probability measures. The proposed algorithm has been tested on several case studies: synthetic data, simulated stirred tank heater system simulation data, and the Tennessee Eastman Process benchmark setup, in the on-line monitoring mode. The performance of the algorithm was evaluated on its ability to detect three types of faults: a sustained bias, an incipient change, and intermittent faults, in contrast with the detection capability of traditional PCA-based metrics such as the Q and Hotelling's T^2 statistics. Furthermore, the performance of the metric was tested for data corrupted by noise from Gaussian as well as non-Gaussian measures. From the results, it was inferred that the proposed PCA model-based fault detection scheme using the KD metric, offers superior performance, as compared to the discussed conventional scheme. The KD metric offers reliable fault detection performance irrespective of the type of noise present in the measurements, and the distribution of the noise itself. Hence, we conclude that the automation of fault detection in practical industrial applications with the KD metric, could prove useful in a variety of areas.

In this chapter, we demonstrated the applicability of the optimal transport distance as a metric for abnormality/fault detection. The process monitoring problem may also be studied under the optimal fault detection system design setting, wherein a fault detection metric, and/or threshold is optimally designed for a system under uncertainty. In Chapter 6, we pivot to this optimal fault detection threshold design problem accounting for ambiguity in the knowledge of a probability distribution describing a multimodal process that is to be monitored for abnormalities; here, we use the formulation presented in Chapter 3 in order to evaluate the worst-case expected performance of a fault detection system for a multimodal process subject to distributional ambiguity.

Chapter 6

Performance Evaluation of a Multimodal Process Fault Detection System using Gaussian Mixture-based Ambiguity Sets

Abstract: Process monitoring of complex, multivariate systems presents a significant challenge in process systems engineering literature. The problem is made further complex when the process is typically operated at multiple operating conditions. In such cases, distinguishing a valid change in operating conditions from abnormal process deviations may be treated as a multimodal process fault detection problem; some methods to address this problem model the process using a multimodal probability distribution. In practical applications, the true distribution, or indeed a good estimate of the same, may not be readily available to the user. Such inexact information on the process' probability distribution lead to poor fault detection performance that may further lead to process operations degradation. To this end, a distributionally robust design of fault detection systems is preferable in the face of ambiguous uncertainty. Recent contributions to process monitoring literature explore the distributionally robust design of fault detection systems that utilize white- or grey-box quantitative models for residual generation and evaluation. In this work, we propose a distributionally robust data-driven approach to fault detection system design that leverages a Bayesian inference-based detection metric in literature for multimodal processes.

6.1 Introduction

Process monitoring forms a significant component in process systems engineering literature, as well as in daily industrial operations to ensure safe and sustainable process operation. Process monitoring, also referred to as abnormal event management (Venkatasubramanian *et al.* 2003c), comprises the main tasks of fault detection, and fault diagnosis, as well as decision-making for mitigation of these faults. The design of fault detection and diagnosis systems may be broadly classified into process model-based (Venkatasubramanian *et al.* 2003c; Venkatasubramanian *et al.* 2003a), and process history-based methods (Venkatasubramanian *et al.* 2003b). While a number of methods from these classes may overlap in their formulations, in essence, process model-based methods use the developed quantitative or qualitative models from process knowledge, while process history-based methods leverage the abundance of data on the process variables to "model" the normal operation. Some examples of quantitative model-based methods include observer-based and parity space-based residual generators, while causal models such as fault trees are classified as qualitative model-based methods. In process history-based methods, a large number of works have used statistical techniques such as principal component analysis (PCA), partial least squares (PLS), and independent component analysis (ICA) to accomplish fault detection (Qin 2012); in particular, a number of variants of these methods have also been proposed for different use cases. A review of work available on process history/data-driven process monitoring is provided in Chapter 5. It may be noted that in recent years, an abundance of machine learning-based techniques have also been leveraged for process monitoring.

The fault detection system may be viewed as a combination of a residual generator, and residual evaluation mechanism (Shang *et al.* 2021). The "residuals" of a process may be generated using a mathematical model, which may be first principles-based or data-driven, describing the normal operating condition(s) and test data from process operation. The residual evaluation step involves designing or choosing an appropriate fault detection metric-threshold combination. Specifically, when the residual evaluation function returns a value larger than a specified threshold, a fault alarm is triggered, and a fault is "detected". In Chapter 4 of this thesis, we addressed the residual evaluation aspect of the fault detection problem through the lens of optimal transport.

The performance of a fault detection (FD) system is heavily influenced by the presence of stochastic disturbances or uncertainties, that may not necessarily be captured by the process model used. The effect of these uncertainties may be observed in practice as a high false alarm rate, wherein data from normal operation incorrectly triggers a fault alarm, or as a reduced fault detection rate, wherein a significant number of abnormal data points are not flagged as faults. Therefore, the optimal design of an FD system for practical use may be posed as an optimization problem under uncertainty, wherein the probability distribution of the uncertainty may be incorporated into the design step (Prékopa 2003). Yet another layer of complexity is added in practical design settings, wherein perfect information is unavailable on the probability distribution of the uncertainty inherent in the process. Therefore, the optimal FD system design under uncertainty is better approached under distributional ambiguity. Such an optimal design problem may be tackled through the lens of distributionally robust optimization (DRO) (Delage and Ye 2010; Wiesemann *et al.* 2014; Abadeh *et al.* 2015; Esfahani and Kuhn 2018). DRO assumes imperfect or "ambiguous" information on the probability distribution of parametric uncertainty in the form of a nominal distribution, and optimizes the problem for the worst-case expected performance of the model over an ambiguity set of candidate distributions constructed around the nominal distribution.

Shang *et al.* (2021) treat the optimal fault detection (FD) system design problem under distributional ambiguity as a distributionally robust chance-constrained problem (DRCCP); here, the authors provide formulations for the optimal design of the FD system with an integrated trade-off between the FAR and FDR performance indices. In their work, Shang *et al.* (2021) aim to maximize the worst-case expected FDR while constraining the worst-case expected FAR in the form of a distributionally robust probabilistic constraint. They presented tractable convex formulations using both moment-based, and the Wasserstein metric-based ambiguity sets. These formulations were built using model-based residual generators from state-space models of the system. It may be noted that this work does not account for FD in multimodal processes. Another work that deals with the optimal FD system design under distributional ambiguity was conduced by Wan *et al.* (2021) that uses a parity relation-based residual generator.

In this work, we consider the fault detection (FD) problem in the context of multimodal processes. A significant challenge in multimodal FD is encountered in distinguishing normal operating condition changes from abnormal operation; this problem is further complicated when distributional ambiguity is accounted for. A number of works in literature have dealt with multimodal FD. In this work, we build upon a Bayesian inference-based probability index developed by Yu and Qin (2008) for multimodal process detection in the context of distributional ambiguity of the process. Other works that have similarly proposed FD schemes (not under distributional ambiguity) include works by Choi *et al.* (2004) and Zhang *et al.* (2021).

In this work, we seek to bridge the gap in a distributionally robust optimal fault detection (FD) system design, using data-driven models developed for multimodal processes. Section 6.2 contains an overview of fault detection theory, and the Bayesian inference-based method chosen from literature (Yu and Qin 2008). Section 6.3 it further presents the proposed optimization models for worst-case performance evaluation of the FD system. Section 6.4 presents an application of the proposed formulations on a synthetic multimodal process case study; here, we demonstrate the effect of ambiguous distributional information on the FD performance, and proceed to quantitatively evaluate the worst-case expected performance based on the FAR and FDR indices. Finally, Section 6.5 summarizes the results of this work, and outlines the future directions.

6.2 Theory

6.2.1 Preliminaries on fault detection

In general process monitoring literature, an anomaly or fault detection (FD) system defined for a process generally comprises two key elements, namely, a residual generator, and a residual evaluator. A residual generator usually comprises known information about the normal operation of a process, generally through first-principles models (such as state-space models), and seeks to give residual information about the data being monitored. When the monitored data is fault-free, the residuals are typically of small magnitude provided that a good process model is used; on the other hand, when the monitored data contains faults, the magnitude of the residuals tends to be large. The residual evaluator is typically given as a fault detection threshold, and is set by the user. When the residual is of a large enough magnitude to exceed this threshold, an alarm is triggered indicating that a fault has been detected. It may be noted that a fault detection system is only as good as the residual generator, and therefore, the process model used. To this end, any uncertainties or ambiguities that have not been accounted for in the design of this residual generator, and/or the detection threshold, may negatively impact the fault detection performance of the system.

In this work, we aim to account for ambiguous information regarding the process model, and evaluate the effect of this ambiguity on fault detection, using a distributionally robust approach, specifically focusing on multimodal processes. In Section 6.2.2, we introduce the residual generator-threshold combination used in this work, and in Section 6.3, we apply the distributionally robust methodology developed in Chapter 3 to evaluate the worst-case expected performance of the fault detection system.

6.2.2 Bayesian inference-based multimodal process fault detection

In this work, we focus on a data-driven approach to fault detection (FD) system design; specifically, we build upon the work of Yu and Qin (2008) wherein the authors propose a Bayesian inference-based probability (BIP) index as a fault detection metric for multimodal processes. In their work, Yu and Qin (2008) model a multimodal process as a Gaussian mixture model (GMM) as follows,

$$\mathbb{P}^0 = \sum_{l=1}^{L} w_l^0 g(\xi|\mu_l^0, \Sigma_l^0)$$

Here, w_l^0 denotes the prior probability or weighting proportion of the l^{th} component in the GMM, while $g(\xi|\mu_l^0, \Sigma_l^0)$ denotes the Gaussian probability density of the vector ξ subject to the l^{th} component. For $m-$dimensional data, the multivariate Gaussian probability density function is given as,

$$g(\xi|\mu_l^0, \Sigma_l^0) = \frac{1}{(2\pi)^{m/2}|\Sigma_l^0|^{1/2}} \exp\Bigg(- \frac{1}{2}(\xi - \mu_l^0)^T(\Sigma_l^0)^{-1}(\xi - \mu_l^0)\Bigg) \tag{6.1}$$

The posterior probability of any monitored sample $\xi_t, \ \forall 1 \leq t \leq N$ may be computed as,

$$p(\theta_l^0|\xi_t) = \frac{w_l^0 g(\xi_t|\mu_l^0, \Sigma_l^0)}{\sum_{l=1}^{L} w_l^0 g(\xi_t|\mu_l^0, \Sigma_l^0)} \tag{6.2}$$

Yu and Qin (2008) define a "local probability index" for each sample $\xi_t, \ \forall 1 \leq t \leq N$, denoted by $P_l^{\text{local}}(\xi_t)$ as,

$$P_{\text{local}}^{(l)}(\xi_t) = \Pr\left\{D_M\left((\xi, \mu_l^0)|\xi \in \mathbb{N}(\mu_l^0, \Sigma_l^0)\right) \leq D_M\left((\xi_t, \mu_l^0)|\xi_t \in \mathbb{N}(\mu_l^0, \Sigma_l^0)\right)\right\} \tag{6.3}$$

Here, $D_M(\xi_t, \mu$ denotes the Mahalanobis distance between ξ_t and μ_l^0 computed as,

$$D_M(\xi_t, \mu_l^0) = (\xi_t - \mu_l^0)^{\mathrm{T}} \Sigma_l^{-1} (\xi_t - \mu_l^0) \tag{6.4}$$

It may be observed that the local probability index $P_{\text{local}}^{(l)}(\xi_t)$, as proposed by Yu and Qin (2008), is an extension to the already-established Mahalanobis distance-based outlier detection scheme. For $m-$dimensional data belonging to a multivariate Gaussian distribution, the Mahalanobis distance between the datapoints and its mean follows a χ_m^2 distribution. In such a case, a datapoint may be treated as an outlier with a confidence level of $(1 - \alpha)$ if its Mahalanobis distance with respect to the population mean is greater than the χ_m^2 cumulative distribution value evaluated at $(1 - \alpha)$. Yu and Qin (2008) extended this idea to that of a multivariate Gaussian mixture distribution, by computing the local probability index for each l^{th} component, which denotes whether or not the datapoint ξ_t is considered an outlier with respect to the l^{th} Gaussian component. The authors further combine these local probability indices $(P_l^{\text{local}}(\xi_t))$ into a single Bayesian inference-based probability (BIP) index using the posterior probabilities of each component as follows,

$$BIP(\xi_t) = \sum_{l=1}^{L} p(\theta_l^0|\xi_t) P_l^{\text{local}}(\xi_t) \tag{6.5}$$

In this FD scheme, a datapoint ξ_t is treated as "faulty" if $BIP(\xi_t) > 1 - \alpha$, where $1 - \alpha$ is the fault detection threshold. It may be further observed that $BIP(\xi_t)$ is directly dependent on the weights of the GMM (w_l^0) through the posterior probability term; therefore, this FD scheme does not account for any ambiguity in the knowledge of the distribution $\mathbb{P}^0$ (more specifically, the weights w_l^0).

In this chapter, we focus on accounting for ambiguity in $\mathbb{P}^0$ as a result of a poor knowledge of the weights w_l^0 by leveraging the distributionally robust optimization

(DRO) framework proposed in Chapter 3 that hedges against an ambiguity set of candidate distributions built around a nominal Gaussian mixture distribution.

6.3 Distributionally robust fault detection system design

In this section, we develop a distributionally robust formulation to evaluate the worst-case performance of the BIP index-based fault detection (FD) system (Yu and Qin 2008) detailed in Section 6.2.2.

We start by considering a "nominal" Gaussian mixture model (GMM) fitted to available multimodal process history data as,

$$\mathbb{P}^0 = \sum_{l=1}^{L} w_l^0 g(\xi|\mu_l^0, \Sigma_l^0)$$

The performance quality of a fault detection system is typically addressed through two criteria, namely, the false alarm rate (FAR) and the fault detection rate (FDR) (Zhang 2016). For a process described by $\mathbb{P}^0$, both FAR and FDR may be defined in a probabilistic form using the BIP index as,

$$FAR = \Pr_{\xi\sim\mathbb{P}^0}\left\{BIP(\xi) > 1-\alpha|f=0\right\} \tag{6.6a}$$

$$FDR = \Pr_{\xi\sim\mathbb{P}^0}\left\{BIP(\xi) > 1-\alpha|f=1\right\} \tag{6.6b}$$

Here, f refers to the "fault label" for a given datapoint, where $f = 0$ denotes "true normal" data and $f = 1$ denotes "true fault" data. Specifically, FAR seeks to find how many normal data points are misclassified as faults, and FDR seeks to find how many faulty data points are correctly classified as faults, using the BIP index for a process described by $\mathbb{P}^0$. A good fault detection system is characterized by a high FDR that also exhibits a low FAR. However, in the face of ambiguity on the knowledge of $\mathbb{P}^0$, it is valuable in practice to find the "worst-case" expected values of FAR and FDR, over an ambiguity set of distributions. This ambiguity set may be defined as a set of candidate distributions constructed around the nominal distribution ($\mathbb{P}^0$) obtained from sampled data; here, we use the W_d metric to characterize the size of this ambiguity set (through the radius ϵ_d). Therefore, the worst-case expected performance problems associated with the FAR and FDR performance indices for a FD system are given as,

$$\max_{\mathbb{P}_\xi \in \mathcal{P}} \; FAR \; := \max_{\mathbb{P}_\xi \in \mathcal{P}} \; \Pr_{\xi \sim \mathbb{P}} \Big\{ BIP(\xi) > 1 - \alpha | f = 0 \Big\} \tag{6.7a}$$

$$\min_{\mathbb{P}_\xi \in \mathcal{P}} \; FDR \; := \min_{\mathbb{P}_\xi \in \mathcal{P}} \; \Pr_{\xi \sim \mathbb{P}} \Big\{ BIP(\xi) > 1 - \alpha | f = 1 \Big\} \tag{6.7b}$$

For a general worst-case expectation maximization of a loss function $\mathcal{L}(\xi)$, the DRO problem using the W_d metric may be written as,

$$\max \quad \mathbb{E}_{\mathbb{P}(\xi)} \Big[\mathcal{L}(\xi) \Big] \tag{6.8a}$$

$$\text{s.t.} \quad W_d(\mathbb{P}, \mathbb{P}^0) \leq \epsilon_d \tag{6.8b}$$

The W_d metric is computed as the square root of the optimal objective value of the optimal transport problem between the nominal GMM ($\mathbb{P}^0$) and a candidate GMM ($\mathbb{P}$) (Chen *et al.* 2018) defined on the same Gaussian components,

$$\mathbb{P}^0 := w_1^0 \mathbb{N}\Big(\mu_1^0, \Sigma_1^0\Big) + w_2^0 \mathbb{N}\Big(\mu_2^0, \Sigma_2^0\Big) + \ldots + w_L^0 \mathbb{N}\Big(\mu_L^0, \Sigma_L^0\Big) \tag{6.9a}$$

$$\mathbb{P} := w_1 \mathbb{N}\Big(\mu_1^0, \Sigma_1^0\Big) + w_2 \mathbb{N}\Big(\mu_2^0, \Sigma_2^0\Big) + \ldots + w_L \mathbb{N}\Big(\mu_L^0, \Sigma_L^0\Big) \tag{6.9b}$$

as,

$$W_d^2(\mathbb{P}^0, \mathbb{P}) := \min_{\pi \in \mathbb{R}^+} \quad \sum_{l=1}^{L} \sum_{l'=1}^{L} \pi_{l,l'} c_{l,l'} \tag{6.10a}$$

$$\text{s.t.} \quad \sum_{l'=1}^{L} \pi_{l,l'} = w_l^0, \quad \forall 1 \leq l \leq L \tag{6.10b}$$

$$\sum_{l=1}^{L} \pi_{l,l'} = w_{l'}, \quad \forall 1 \leq l' \leq L \tag{6.10c}$$

Here, $c_{l,l'}$ refers to the cost of transport between the Gaussian components $\mathbb{N}\Big(\mu_l^0, \Sigma_l^0\Big)$ and $\mathbb{N}\Big(\mu_{l'}^0, \Sigma_{l'}^0\Big)$, and is computed using the closed-form expression for the squared 2-Wasserstein distance (Takatsu 2011) as,

$$c_{l,l'} := ||\mu_l^0 - \mu_{l'}^0||^2 + \text{Tr}\Bigg[\Sigma_l^0 + \Sigma_{l'}^0 - 2 \Bigg(\Big[\Sigma_l^0\Big]^{\frac{1}{2}} \Sigma_{l'}^0 \Big[\Sigma_l^0\Big]^{\frac{1}{2}} \Bigg)^{\frac{1}{2}} \Bigg] \tag{6.11}$$

Then, the worst-case expectation maximization of $\mathcal{L}(\xi)$ using the W_d ambiguity set may be written as,

$$\begin{aligned}
\max \quad & \mathbb{E}_{\mathbb{P}(\xi)}\big[\mathcal{L}(\xi)\big] && (6.12a)\\
\text{s.t.} \quad & \min_{\pi\in\mathbb{R}^+} \quad \sum_{l=1}^{L}\sum_{l'=1}^{L} \pi_{l,l'}c_{l,l'} \le \epsilon_d^2 && (6.12b)\\
& \qquad \text{s.t.} \quad \sum_{l'=1}^{L} \pi_{l,l'} = w_l^0, \quad \forall 1 \le l \le L && (6.12c)\\
& \qquad \qquad \sum_{l=1}^{L} \pi_{l,l'} = w_{l'}, \quad \forall 1 \le l' \le L && (6.12d)
\end{aligned}$$

The "min" operator in Constraint 6.12b may be neglected. When $\mathcal{L}(\xi)$ is replaced by the probabilistic formulations of FAR and FDR in Equations 6.6a and 6.6b, respectively, the worst-case FAR and FDR formulations are obtained. For computational tractability, we choose to replace the probabilistic function by its equivalent expectation operator taken over the indicator function as,

$$\mathrm{Pr}_{\xi\sim\mathbb{P}^0}\big\{\mathcal{L}(\xi)\big\} \quad := \quad \mathbb{E}_{\mathbb{P}}\big[\mathbb{I}\big(\mathcal{L}(\xi)\big)\big] \qquad (6.13)$$

The expectation operator may be empirically estimated for $\xi_t, \quad 1 \le t \le N$, with corresponding fault labels (f_t), wherein $f_t = 0$ indicates a normal point, and $f_t = 1$ indicates a faulty point. Then, the empirical estimates for the expected values of FAR and FDR are given as,

$$\mathbb{E}_{\mathbb{P}}\big[FAR\big] = \mathbb{E}_{\mathbb{P}}\big[\mathbb{I}\big(BIP(\xi) > 1-\alpha | f = 0\big)\big] := \frac{\sum_{t=1}^{N}((1-f_t)\mathbb{I}\big(BIP(\xi_t) > 1-\alpha\big)}{\sum_{t=1}^{N} 1 - f_t} \qquad (6.14a)$$

$$\mathbb{E}_{\mathbb{P}}\big[FDR\big] = \mathbb{E}_{\mathbb{P}}\big[\mathbb{I}\big(BIP(\xi) > 1-\alpha | f = 1\big)\big] := \frac{\sum_{t=1}^{N}((f_t)\mathbb{I}\big(BIP(\xi_t) > 1-\alpha\big)}{\sum_{t=1}^{N} f_t} \qquad (6.14b)$$

Finally, the worst-case expected FAR problem is given as,

$$\begin{aligned}
\max_{w,\pi\in\mathbb{R}^+} \quad & \frac{\sum_{t=1}^{N}((1-f_t)\mathbb{I}\big(BIP(\xi_t) > 1-\alpha\big)}{\sum_{t=1}^{N} 1 - f_t} && (6.15a)\\
\text{s.t.} \quad & \sum_{l=1}^{L}\sum_{l'=1}^{L} \pi_{l,l'}c_{l,l'} \le \epsilon_d^2 && (6.15b)
\end{aligned}$$

$$\sum_{l'=1}^{L} \pi_{l,l'} = w_l^0, \quad \forall 1 \le l \le L \tag{6.15c}$$

$$\sum_{l=1}^{L} \pi_{l,l'} = w_{l'}, \quad \forall 1 \le l' \le L \tag{6.15d}$$

$$BIP(\xi_t) = \sum_{l=1}^{L} p(\theta_l^0|\xi_t) P_l^{\text{local}}(\xi_t), \quad \forall 1 \le t \le N \tag{6.15e}$$

$$p(\theta_l^0|\xi_t) = \frac{w_l^0 g(\xi_t|\mu_l^0, \Sigma_l^0)}{\sum_{l=1}^{L} w_l^0 g(\xi_t|\mu_l^0, \Sigma_l^0)}, \quad \forall 1 \le t \le N \tag{6.15f}$$

$$g(\xi_t|\mu_l^0, \Sigma_l^0) = \frac{1}{(2\pi)^{m/2}|\Sigma_l^0|^{1/2}} \exp\Bigg(-\frac{1}{2}(\xi_t - \mu_l^0)^T (\Sigma_l^0)^{-1} (\xi_t - \mu_l^0)\Bigg), \quad \forall 1 \le t \le N \tag{6.15g}$$

$$P_l^{\text{local}}(\xi_t) = \mathit{chi2cdf}\left(D_M(\xi_t, \mu_l^0), m\right), \quad \forall 1 \le t \le N, \quad 1 \le l \le L \tag{6.15h}$$

and the worst-case expected FDR problem as,

$$\min_{w,\pi \in \mathbb{R}^+} \quad \frac{\sum_{t=1}^{N} (f_t \mathbb{I}\big(BIP(\xi_t) > 1 - \alpha\big)}{\sum_{t=1}^{N} f_t} \tag{6.16a}$$

$$\text{s.t.} \quad \sum_{l=1}^{L} \sum_{l'=1}^{L} \pi_{l,l'} c_{l,l'} \le \epsilon_d^2 \tag{6.16b}$$

$$\sum_{l'=1}^{L} \pi_{l,l'} = w_l^0, \quad \forall 1 \le l \le L \tag{6.16c}$$

$$\sum_{l=1}^{L} \pi_{l,l'} = w_{l'}, \quad \forall 1 \le l' \le L \tag{6.16d}$$

$$BIP(\xi_t) = \sum_{l=1}^{L} p(\theta_l^0|\xi_t) P_l^{\text{local}}(\xi_t), \quad \forall 1 \le t \le N \tag{6.16e}$$

$$p(\theta_l^0|\xi_t) = \frac{w_l^0 g(\xi_t|\mu_l^0, \Sigma_l^0)}{\sum_{l=1}^{L} w_l^0 g(\xi_t|\mu_l^0, \Sigma_l^0)}, \quad \forall 1 \le t \le N \tag{6.16f}$$

$$g(\xi_t|\mu_l^0, \Sigma_l^0) = \frac{1}{(2\pi)^{m/2}|\Sigma_l^0|^{1/2}} \exp\Bigg(-\frac{1}{2}(\xi_t - \mu_l^0)^T (\Sigma_l^0)^{-1} (\xi_t - \mu_l^0)\Bigg), \quad \forall 1 \le t \le N \tag{6.16g}$$

$$P_l^{\text{local}}(\xi_t) = \mathit{chi2cdf}\left(D_M(\xi_t, \mu_l^0), m\right), \quad \forall 1 \le t \le N, \quad 1 \le l \le L \tag{6.16h}$$

Here, m refers to the dimensionality of the multimodal process. It may be noted that the presence of the indicator function (taken over a nonlinear function of w) in the objective functions of Models 6.15 and 6.16 gives rise to a non-convex formulation. In this work, we solve these models using the *surrogateopt* heuristic solver in MATLAB (Wang and Shoemaker 2014; Regis and Shoemaker 2007).

6.4 Worst-case performance evaluation of a multimodal process fault detection system - a case study

In this section, we illustrate how the worst-case performance of a fault detection (FD) system may be evaluated under distributional ambiguity regarding the Gaussian mixture distribution that describes a multimodal process on a synthetic case study. Section 6.4.1 contains an overview of the dataset generation process, while Section 6.4.2 discusses the effect that ambiguity of the process distribution has over the fault detection performance using the BIP criterion (Yu and Qin 2008). Sections 6.4.3 and 6.4.4 illustrate the worst-case performance criteria of the FD system, namely the false alarm rate (FAR) and fault detection rate (FDR), respectively.

6.4.1 Dataset generation

In this study, we generated a synthetic dataset to simulate a two-dimensional, three-operating-mode process from the (true) Gaussian mixture distribution,

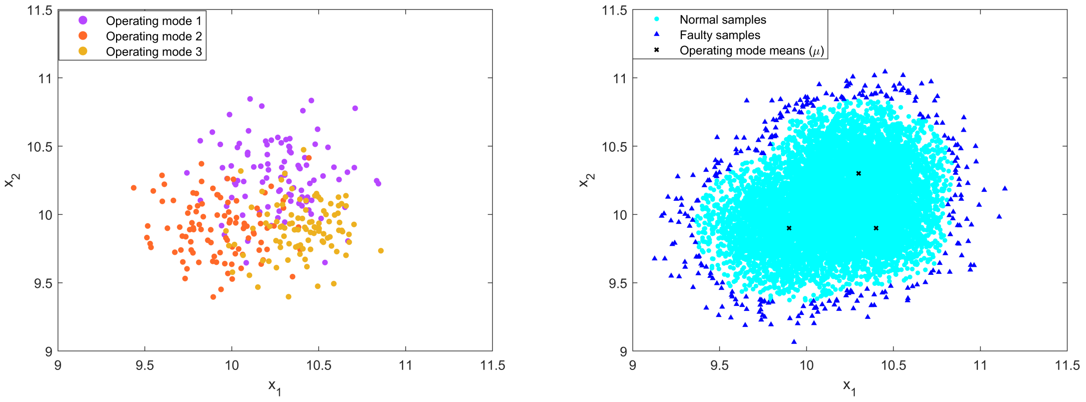
$$\mathbb{P}^{\text{true}} := 0.45\mathbb{N}(\mu_1, \Sigma_1) + 0.35\mathbb{N}(\mu_2, \Sigma_2) + 0.2\mathbb{N}(\mu_3, \Sigma_3) \tag{6.17}$$

where,

$$\mu_1 = \begin{bmatrix} 10.3 \\ 10.3 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 0.05 & 0 \\ 0 & 0.05 \end{bmatrix}, \mu_2 = \begin{bmatrix} 9.9 \\ 9.9 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 0.05 & 0 \\ 0 & 0.05 \end{bmatrix}, \mu_3 = \begin{bmatrix} 10.4 \\ 9.9 \end{bmatrix},$$

$$\Sigma_3 = \begin{bmatrix} 0.03 & 0 \\ 0 & 0.03 \end{bmatrix}$$

The "process" data that is generated by the GMM in 6.17 is illustrated in Figure 6.1a; from this figure, it is evident that the data points are not well-separated into distinct clusters, and there is significant overlap of data from different operating modes.



(a) (b)

Figure 6.1: (a) A single dataset of sampled data from the synthetic three-operating-mode process described by $\mathbb{P}^{\text{true}}$ in 6.17, (b) Classification of points generated from $\mathbb{P}^{\text{true}}$ into "true normal" and "true fault" datapoints

In order to evaluate the fault detection (FD) performance of the BIP index-based scheme, as proposed by Yu and Qin (2008), we use the following performance metrics - false alarm rate (FAR), and fault detection rate (FDR) - as described in Section 6.3. The evaluation of FAR requires a set of "true normal" data in order to evaluate how many datapoints are being misclassified as faults, while the evaluation of FDR requires a set of "true fault" data to evaluate their misclassification as datapoints from normal operation. To this end, we generated a set of 10000 datapoints from $\mathbb{P}^{\text{true}}$ and classified all points whose Mahalanobis distance from any of the three component means (μ_1, μ_2, μ_3) was larger than that of the inverse χ^2_2 distribution value for a confidence interval of 95% as "true faults", and the rest as "true normal". These datasets are illustrated in Figure 6.1b. It may be noted that, out of 10000 points generated from $\mathbb{P}^{\text{true}}$, 9721 points were classified as true normal data, and the remaining 279 as true fault data.

6.4.2 Effect of inexactness of probability distribution information on performance

The fault detection (FD) scheme put forth by Yu and Qin (2008) models a multimodal process as a Gaussian mixture model (GMM). From the scheme explained in Section 6.2.2, it may be seen that the FD metric (BIP index) is affected by the posterior probability of the components in this GMM, which is further affected by the prior probabilities of the GMM components. Therefore, the efficacy of the BIP index as an FD metric may be seen as significantly dependent on the quality of information available on the GMM. In practical applications, the GMM is obtained from process history data, and thus, the accuracy of the fitted GMM is sensitive to sampling. To illustrate the effect of the fitted GMM ($\mathbb{P}^0$) on the FD performance, we generated 100 trial datasets from $\mathbb{P}^{\text{true}}$, and evaluated the false alarm rate (FAR) and fault detection rate (FDR) using the BIP index. Here, we calculated FAR and FDR using the indicator function-based formulations presented in Equations 6.14a - 6.14b.

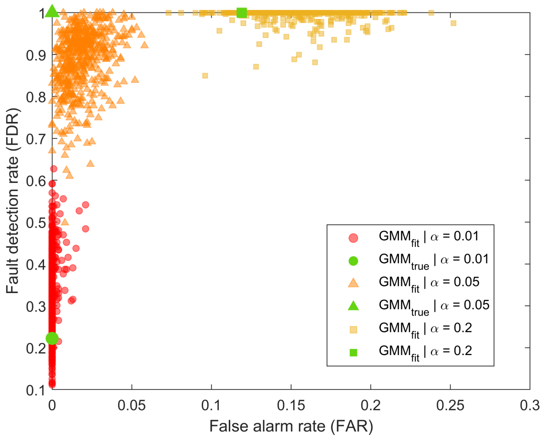


Figure 6.2: An illustration of the effect of ambiguity in the multimodal process distribution knowledge on fault detection performance for $\alpha = [0.01, 0.05, 0.2]$.

Figure 6.2 highlights the effect of ambiguity of knowledge about $\mathbb{P}^{\text{true}}$ that is inherent when $\mathbb{P}^0$ is obtained from process history data. It is evident that the performance of the BIP index-based FD system is significantly affected by goodness-of-fit of $\mathbb{P}^0$

to the sampled data. This result motivates the need for quantifying the worst-case performance that may be expected from an FD system using a fitted GMM ($\mathbb{P}^0$), for different levels of ambiguity (quantified by the ambiguity set size/radius). In addition, this figure also provides context for the need to choose the fault detection threshold (α) carefully. We observe that the performance of the BIP index-based FD scheme is significantly affected by α even for perfect information ($\mathbb{P}^{\text{true}}$) on the process distribution.

6.4.3 Worst-case FAR performance evaluation using distributionally robust optimization

Having illustrated the motivation for evaluating the worst-case performance of the BIP index-based fault detection (FD) system (Yu and Qin 2008) in Section 6.4.2, we now consider the FD problem under distributional ambiguity (the DRO-FD problem). Specifically, we assume ambiguity in the knowledge of the weights of the components of the Gaussian mixture model (GMM) fitted to a sample dataset of 300 points from $\mathbb{P}^{\text{true}}$. In this section, we aim to quantify the worst-case FD performance using the false alarm rate (FAR) criterion under the DRO setting, and solve for the worst-case expected FAR using Model 6.15.

Figure 6.3 illustrates the evolution of the worst-case FAR with increasing levels of ambiguity, characterized by the W_d ambiguity set radius (ϵ_d), for various fault detection thresholds ($1-\alpha$). We see that, as we consider increasing levels of ambiguity regarding the weighting proportions of the fitted GMM ($\mathbb{P}^0$), the worst-case FAR also increases. Furthermore, we see that for a specific ϵ_d, the worst-case FAR is higher for larger α (that is, for a lower fault detection threshold).

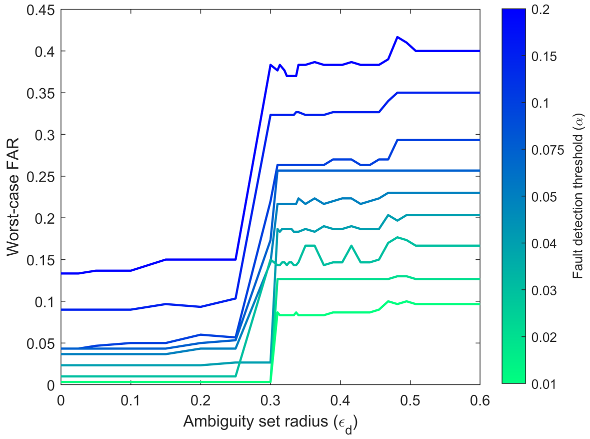
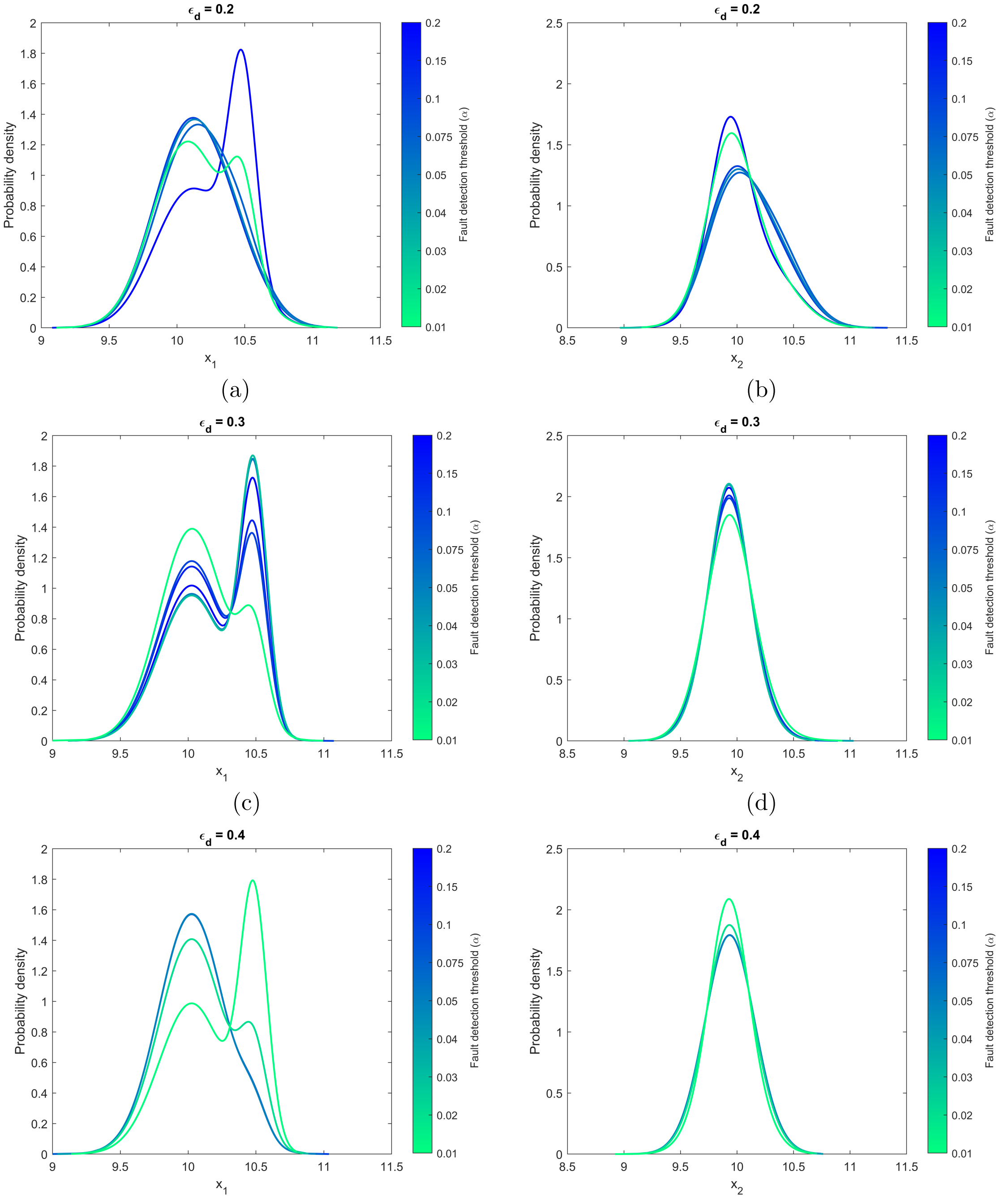


Figure 6.3: Evolution of the worst-case false alarm rate (FAR) with increasing ambiguity, characterized by ϵ_d, for various fault detection thresholds $(1 - \alpha)$

The worst-case FAR for a given ambiguity set radius (ϵ_d) is obtained when the optimization problem in Model 6.15 hedges against that candidate distribution, denoted by $\mathbb{P}^{\mathrm{wc}}$ that gives the maximum value. Figure 6.4 illustrates some worst-case distributions; Figures 6.4a - 6.4b depict the $\mathbb{P}^{wc}$ for $\epsilon_d = 0.2$ for x_1 and x_2, respectively, while Figures 6.4c - 6.4d, and Figures 6.4e - 6.4f depict $\mathbb{P}^{\mathrm{wc}}_{\epsilon_d=0.3}$, and $\mathbb{P}^{\mathrm{wc}}_{\epsilon_d=0.4}$, respectively.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 6.4: An illustration of the worst-case GMMs, for various α, associated with the worst-case FAR problem for $\epsilon_d = 0.2$ [(a) - (b)], $\epsilon_d = 0.3$ [(c) - (d)], and $\epsilon_d = 0.4$ [(e) - (f)]

6.4.4 Worst-case FDR performance evaluation using distributionally robust optimization

Having illustrated the worst-case performance of the BIP index-based fault detection (FD) system under distributional ambiguity, using the false alarm rate (FAR) criterion in Section 6.4.3, we now move on to an illustration of the same using the fault detection rate (FDR) criterion. In contrast to the worst-case FAR problem, wherein the goal is to evaluate what the maximum value of FAR is for a given ambiguity level, in the worst-case FDR problem, the goal is to evaluate the minimum value of FDR. We solved for the worst-case expected FDR using Model 6.16.

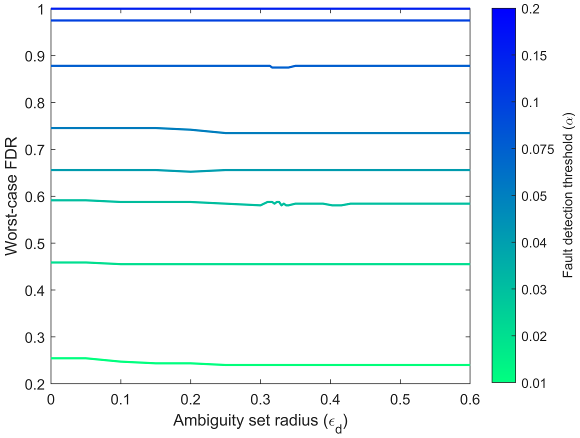
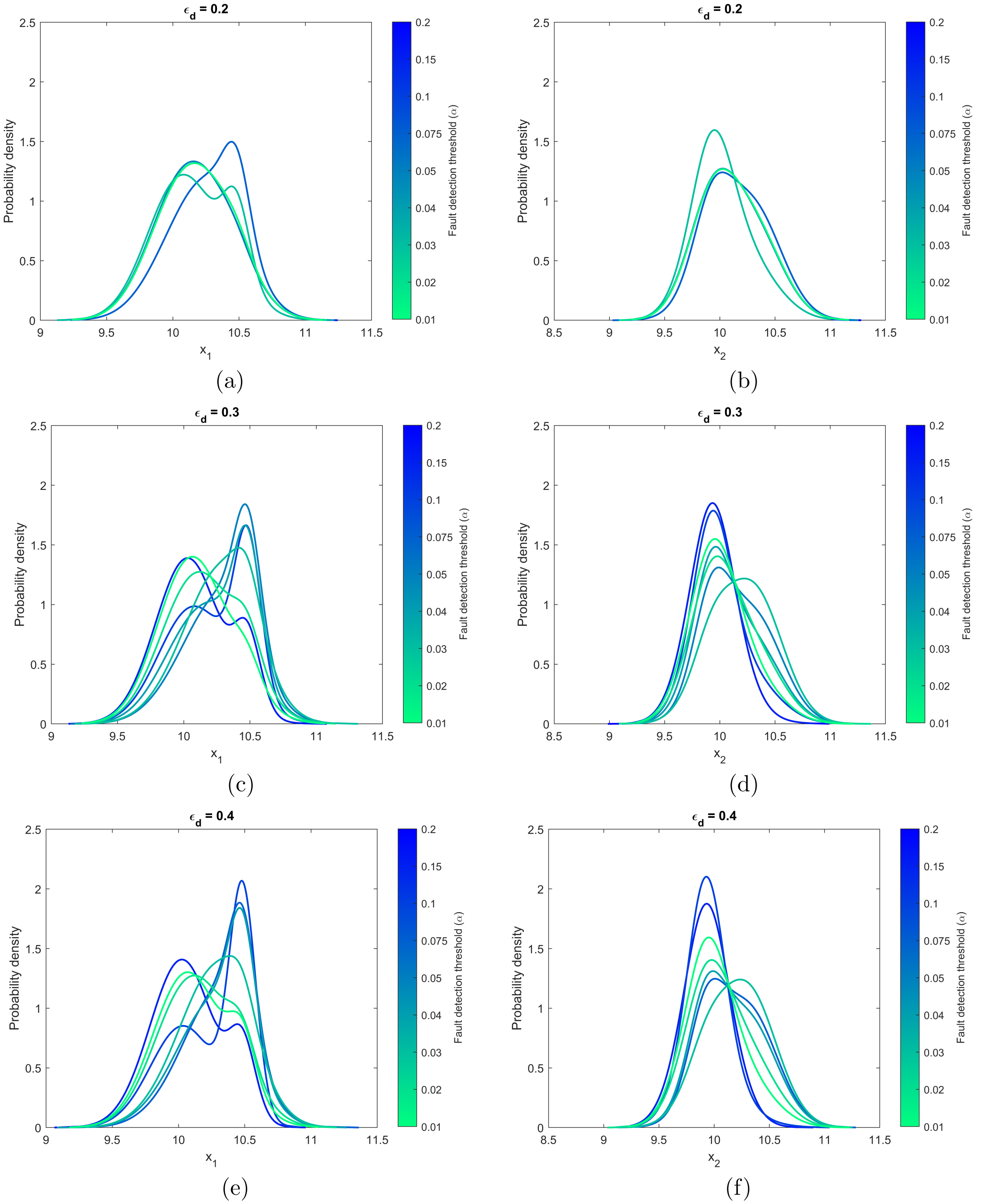


Figure 6.5: Evolution of the worst-case fault detection rate (FDR) with increasing ambiguity, characterized by ϵ_d, for various fault detection thresholds $(1-\alpha)$

Figure 6.5 illustrates the evolution of the worst-case FDR for increasing ϵ_d, for various fault detection thresholds $(1-\alpha)$. We see that, for increasing levels of ambiguity regarding the fitted $\mathbb{P}^0$, the worst-case FDR shows little to no variation. Furthermore, for a specific ϵ_d, the worst-case FDR is higher for larger α (that is, for lower detection thresholds). In contrast to the worst-case FAR results, we see that the worst-case FDR is less sensitive to the level of ambiguity surrounding the fitted $\mathbb{P}^0$, and that the FDR is generally more dependent on the fault detection threshold $(1-\alpha)$.



(a) (b) (c) (d) (e) (f)

Figure 6.6: An illustration of the worst-case GMMs, for various α, associated with the worst-case FDR problem for $\epsilon_d = 0.2$ [(a) - (b)], $\epsilon_d = 0.3$ [(c) - (d)], and $\epsilon_d = 0.4$ [(e) - (f)]

Figure 6.6 illustrates worst-case distributions ($\mathbb{P}^{\text{wc}}$) for the ambiguity set radii of $\epsilon_d = [0.2, 0.3, 0.4]$. Comparing with the $\mathbb{P}^{\text{wc}}$ obtained for the same ϵ_d values for the worst-case FAR problem, we see that the $\mathbb{P}^{\text{wc}}$ associated with the worst-case FDR problem are markedly different. It may be noted that the worst-case distribution, as defined in the distributionally robust optimization setting, is dependent not only on the ambiguity set radius, but also on the objective function being optimized. Since the objectives of the worst-case FAR and worst-case FDR problems are different, it is expected to obtain different $\mathbb{P}^{\text{wc}}$ in each case.

6.5 Summary and future directions

In this work, we aimed to evaluate the worst-case fault detection (FD) performance for a multimodal process under distributional ambiguity. We chose a data-driven multimodal process fault detection method available in literature that treats the process as a Gaussian mixture, and derived distributionally robust formulations to evaluate the worst case expected FD performance in terms of two indices, namely the false alarm rate (FAR) and the fault detection rate (FDR). We developed these formulations using the W_d DRO method proposed in Chapter 3, that uses optimal transport between Gaussian mixtures to construct the ambiguity set involved in the problem. We demonstrated the effect of ambiguity on the FD performance, and illustrated the use of these models on a synthetic case study, wherein we studied the worst-case expected FD performance for a 3-operating-mode process, as well as the worst-case distributions corresponding to different levels of ambiguity and fault detection thresholds. In the future, we aim to extend this study to a worst-case expected (combined) FAR-FDR performance study, that will enable the user to choose the appropriate fault detection threshold for their choice of FAR-FDR, and specified level of ambiguity regarding the process distribution. We also aim to demonstrate the performance of the proposed methodology on simulation case studies and benchmark datasets to generalize applicability.

Chapter 7

Summary and Future Directions

This thesis presents contributions to the process systems engineering literature in the areas of process optimization and process monitoring. The proposed algorithms, formulations, and mathematical studies conducted are unified by the common theme of optimal transport theory. Under process optimization, this thesis presents contributions to the fields of stochastic optimization (Chapter 2), and distributionally robust optimization (Chapters 3 and 4). In the area of process monitoring, works and studies presented in this thesis contribute to the areas of on-line change-point and fault detection (Chapter 5), as well as optimal design of process monitoring systems under uncertainty (Chapter 6). This chapter summarizes the problem statements tackled in this thesis, the knowledge gaps addressed by each body of work, and potential directions for future contributions.

7.1 Improvements to stochastic programming

Chapter 2 of this thesis focuses on the field of scenario-based stochastic programming, which is a well-established and well-utilized method in optimization to solve problems under uncertainty. Under this setting, the parametric uncertainty present in the problem is approximated using discrete realizations, also termed as scenarios. The resulting solution quality is found to be dependent significantly on the quality of scenarios chosen that well approximate the true probability distribution, as well as the number of scenarios considered. However, performing stochastic optimization with a large scenario superset proves computationally difficult, and therefore, the task of optimally reducing this superset to a smaller subset while preserving information is a key task explored in optimization literature.

This thesis proposes a method for optimal scenario reduction using an entropy-regularized variant of the optimal transport problem. While the conventional optimal transport problem has been used in literature previously to accomplish this task, large-dimensional supersets are known to cause memory bottlenecks in linear programming. In such cases, the availability of an analytical solution to the entropy-regularized variant offers an advantageous numerical iterative scheme through the use of the Sinkhorn-Knopp algorithm (Sinkhorn 1964; Sinkhorn 1967; Sinkhorn and Knopp 1967; Cuturi 2013) which this thesis leverages for optimal scenario reduction. This algorithm was further extended in order to obtain multistage scenario trees from large scenario fans for multistage stochastic optimization. The use of these algorithms was demonstrated on two case studies, namely a two-stage problem, and a multi-stage problem. In both cases, it was seen that the use of entropy-regularized optimal transport for scenario reduction decreases the computational time while providing solutions with good accuracy with respect to those obtained using the original superset. It may be noted that this work considers the stochastic programming “input”-matching problem, which seeks to generate a smaller subset of scenarios from a large superset whose probability distributions are as similar as possible. One possible future direction for this work would be to consider the “performance”-matching problem, wherein an optimal subset is generated whose worst-case expected performance matches that of the superset. Another future direction may be to address the scenario reduction problem wherein the data contains a mix of continuous, as well as categorical dimensions.

7.2 Novel formulations for distributionally robust optimization

Chapters 3 and 4 of this thesis contribute to the field of distributionally robust optimization. This method of optimization under uncertainty retains favorable aspects of both stochastic and robust optimization frameworks, and gives solutions robust to distributional ambiguity. Chapter 3 proposes a mathematical formulation for the worst-case expectation maximization problem associated with the distributionally robust optimization problem. This work assumed that the probability distribution of the multimodal uncertainty inherent in a problem may be modeled as a Gaussian mixture, whose probability distribution is ambiguous. Subsequently, an ambiguity set containing candidate distributions was built around a nominal (estimated) distribution using an optimal transport metric for Gaussian mixtures (Chen *et al.* 2018). Therefore, this work combines elements of both moment-based, as well as metric-based ambiguity set construction methods, by retaining first- and second-order moments observed in the sampled data in an optimal transport metric-based setting. The efficacy of the proposed model was demonstrated on an illustrative case study, as well as a financial portfolio optimization study. Chapter 4 extends this proposed model to a distributionally robust chance-constrained setting, wherein the worst-case expected violation of constraints under the ambiguity set was constrained to a user-defined threshold. The use of this tractable formulation was demonstrated on two case studies, a blending problem, as well as a chemical process design study. In both chapters, the performance of the proposed formulations were compared to that of the conventional Wasserstein distance-based approach, and it was seen that the proposed formulation exhibits superior performance. Furthermore, a study on the worst-case distribution, that is, the extremal distribution for a given ambiguity set radius against which the solution hedges, is presented in both chapters. Through this study, Chapter 3 also presents a method to calculate an upper bound on the radius of the ambiguity set, which is a hyperparameter that significantly affects the performance of a distributionally robust optimization problem.

In the formulations proposed in Chapters 3 and 4, it is assumed that uncertainty involved in the optimization problem exhibits multiple modes. It is further noted that the efficiency of Gaussian mixture models in approximating arbitrary distributions has been reported in literature. One possible future direction for this work would be to extend the formulations to generic mixture models for which an optimal transport-type metric has been recently proposed (Dusson *et al.* 2023).

7.3 Optimal transport for process monitoring

Chapters 5 and 6 of this thesis apply the optimal transport problem to problems in process monitoring. This thesis tackles the fault detection problem under two settings. Chapter 5 explores the use of the optimal transport distance as a metric for change-point detection, and fault detection. In this work, a moving window scheme was utilized on residual signals from a data-driven principal component analysis model, and solve the optimal transport problem between signals to find the level of similarity between them. The efficiency of the optimal transport distance was demonstrated in identifying changes to process operations, as well as detecting faults relative to normal process behavior even for smaller magnitude faults. The use of the proposed fault detection algorithm was illustrated on a synthetic dataset, a simulated stirred tank heater setup, as well as on the benchmark Tennessee Eastman process. In all three cases, the fault detection capability of the optimal transport distance was found to be superior to that of conventional PCA indices such as the squared prediction error (or the Q statistic), and the Hotelling's T^2 statistic.

Chapter 6 of this thesis approaches the fault detection problem through the lens of optimal design. This work considered the effect of uncertainty on fault detection performance of a system through distributional ambiguity, specifically for processes operating at multiple modes. In this context, the existing literature gap on distributionally robust fault detection system design is bridged in two ways. Firstly, a formulation for multimodal processes was developed, while a majority of current literature does not focus on the same. Secondly, a data-driven approach was applied for fault detection system design under ambiguity; currently, the literature on this topic utilizes first principle-based models, such as state space models (Shang *et al.* 2021), to address this issue. This work considered a Bayesian inference-based index for multimodal process fault detection proposed in literature that models the process as a Gaussian mixture (Yu and Qin 2008), and extended the same accounting for uncertainty in the fitting of this mixture model to process history data. To this end, the distributionally robust optimization framework proposed in Chapter 3 was applied in the fault detection framework to evaluate the worst-case performance of a fault detection system, using the indices of fault detection rate, and false alarm rate. The worst-case expected performance models pertaining to these indices were formulated, and their use was demonstrated on a synthetic multimodal case study. The evolution of the worst-case expected performance of a fault detection system was tracked for varying levels of ambiguity quantified by the ambiguity set radius. Using these results, recommendations may be made for an appropriate choice of the fault detection threshold magnitude for acceptable worst-case performance under a defined ambiguity level.

The worst-case expected performance models in Chapter 6 were formulated separately for the fault detection rate, and false alarm rate. A possible future direction for this work is to combine both indices and solve for a unified worst-case evaluation of the fault detection system performance under distributional ambiguity of the multimodal process. Additional experiments on benchmarks setups and process data would also be beneficial to assess the applicability of the proposed formulations for practical decision-making under uncertainty.

Bibliography

[1] T. F. Edgar, D. M. Himmelblau, and L. S. Lasdon, “Optimization of chemical processes,” *(No Title)*, 2001.

[2] N. V. Sahinidis, “Optimization under uncertainty: State-of-the-art and opportunities,” *Computers & chemical engineering*, vol. 28, pp. 971–983, 6-7 2004.

[3] C. Ning and F. You, “Optimization under uncertainty in the era of big data and deep learning: When machine learning meets mathematical programming,” *Computers & Chemical Engineering*, vol. 125, pp. 434–448, 2019.

[4] A. J. Keith and D. K. Ahner, “A survey of decision making and optimization under uncertainty,” *Annals of Operations Research*, vol. 300, pp. 319–353, 2 2021.

[5] R. Isermann, “Process fault detection based on modeling and estimation methods—a survey,” *automatica*, vol. 20, pp. 387–404, 4 1984.

[6] B. M. Wise and N. B. Gallagher, “The process chemometrics approach to process monitoring and fault detection,” *Journal of Process Control*, vol. 6, pp. 329–348, 6 1996.

[7] S. J. Qin, “Survey on data-driven industrial process monitoring and diagnosis,” *Annual reviews in control*, vol. 36, pp. 220–234, 2 2012.

[8] Z. Ge, Z. Song, and F. Gao, “Review of recent research on data-based process monitoring,” *Industrial & Engineering Chemistry Research*, vol. 52, pp. 3543–3562, 10 2013.

[9] K. Severson, P. Chaiwatanodom, and R. D. Braatz, “Perspectives on process monitoring of industrial systems,” *Annual Reviews in Control*, vol. 42, pp. 190–200, 2016.

[10] G. B. Dantzig, “Linear programming under uncertainty,” *Management science*, vol. 1, pp. 197–206, 3-4 1955.

[11] E. M. L. Beale, “On minimizing a convex function subject to linear inequalities,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 17, pp. 173–184, 2 1955.

[12] A. Shapiro and A. Philpott, “A tutorial on stochastic programming,” *Manuscript. Available at www2. isye. gatech. edu/ashapiro/publications. html*, vol. 17, 2007.

[13] S. Ahmed and A. Shapiro, *The sample average approximation method for stochastic programs with integer recourse. optimization online*, 2002.

[14] J. Luedtke and S. Ahmed, "A sample approximation approach for optimization with probabilistic constraints," *SIAM Journal on Optimization*, vol. 19, pp. 674–699, 2 2008.

[15] G. B. Dantzig and G. Infanger, "Large-scale stochastic linear programs: Importance sampling and benders decomposition," vol. 91, 1991, p. 111.

[16] M. A. H. Dempster and R. T. Thompson, "Evpi-based importance sampling solution proceduresfor multistage stochastic linear programmeson parallel mimd architectures," *Annals of Operations Research*, vol. 90, pp. 161–184, 0 1999.

[17] K. Høyland and S. W. Wallace, "Generating scenario trees for multistage decision problems," *Management science*, vol. 47, pp. 295–307, 2 2001.

[18] G. C. Pflug, "Scenario tree generation for multiperiod financial optimization by optimal discretization," *Mathematical programming*, vol. 89, pp. 251–271, 2001.

[19] K. Høyland, M. Kaut, and S. W. Wallace, "A heuristic for moment-matching scenario generation," *Computational optimization and applications*, vol. 24, pp. 169–185, 2003.

[20] H. Heitsch and W. Römisch, "Scenario tree reduction for multistage stochastic programs," *Computational Management Science*, vol. 6, pp. 117–133, 2 2009.

[21] H. Heitsch and W. Römisch, "Scenario reduction algorithms in stochastic programming," *Computational optimization and applications*, vol. 24, pp. 187–206, 2003.

[22] J. Dupačová, N. Gröwe-Kuska, and W. Römisch, "Scenario reduction in stochastic programming," *Mathematical programming*, vol. 95, pp. 493–511, 3 2003.

[23] A. Charnes and W. W. Cooper, "Chance-constrained programming," *Management science*, vol. 6, pp. 73–79, 1 1959.

[24] B. L. Miller and H. M. Wagner, "Chance constrained programming with joint constraints," *Operations Research*, vol. 13, pp. 930–945, 6 1965.

[25] A. Ben-Tal and A. Nemirovski, "Robust solutions of linear programming problems contaminated with uncertain data," *Mathematical programming*, vol. 88, pp. 411–424, 2000.

[26] R. T. Rockafellar, S. Uryasev, *et al.*, "Optimization of conditional value-at-risk," *Journal of risk*, vol. 2, pp. 21–42, 2000.

[27] A. Nemirovski and A. Shapiro, "Convex approximations of chance constrained programs," *SIAM Journal on Optimization*, vol. 17, pp. 969–996, 4 2007.

[28] A. L. Soyster, "Convex programming with set-inclusive constraints and applications to inexact linear programming," *Operations research*, vol. 21, pp. 1154–1157, 5 1973.

[29] A. Ben-Tal, A. Goryashko, E. Guslitzer, and A. Nemirovski, "Adjustable robust solutions of uncertain linear programs," *Mathematical programming*, vol. 99, pp. 351–376, 2 2004.

[30] O. Boni and A. Ben-Tal, “Adjustable robust counterpart of conic quadratic problems,” *Mathematical Methods of Operations Research*, vol. 68, pp. 211–233, 2008.

[31] İ. Yanıkoğlu, B. L. Gorissen, and D. den Hertog, “A survey of adjustable robust optimization,” *European Journal of Operational Research*, vol. 277, no. 3, pp. 799–813, 2019.

[32] D. Bertsimas and M. Sim, “The price of robustness,” *Operations research*, vol. 52, pp. 35–53, 1 2004.

[33] J. M. Keynes, *A Treatise on Probability.* Dover Publications, 1921.

[34] F. H. Knight, *Risk, uncertainty and profit.* Houghton Mifflin, 1921, vol. 31.

[35] M. D. Packard, P. L. Bylund, and B. B. Clark, “Keynes and knight on uncertainty: Peas in a pod or chalk and cheese?” *Cambridge Journal of Economics*, vol. 45, pp. 1099–1125, 5 2021.

[36] K. J. Arrow, “Alternative approaches to the theory of choice in risk-taking situations,” *Econometrica: Journal of the Econometric Society*, pp. 404–437, 1951.

[37] S Karlin, K Arrow, and H Scarf, “A min–max solution of an inventory problem,” *Studies in the international theory of inventory and productions. Stanford University Press, Stanford*, 1958.

[38] J. Žáčková, “On minimax solutions of stochastic linear programming problems,” *Časopis pro pěstování matematiky*, vol. 91, no. 4, pp. 423–430, 1966.

[39] G. Bayraksan and D. K. Love, “Data-driven stochastic programming using phi-divergences,” in INFORMS, 2015, pp. 1–19.

[40] K. Postek, D. den Hertog, and B. Melenberg, “Computationally tractable counterparts of distributionally robust constraints on risk measures,” *SIAM Review*, vol. 58, pp. 603–650, 4 2016.

[41] A. Shapiro, “Tutorial on risk neutral, distributionally robust and risk averse multistage stochastic programming,” *European Journal of Operational Research*, vol. 288, pp. 1–13, 1 2021.

[42] F. Lin, X. Fang, and Z. Gao, “Distributionally robust optimization: A review on theory and applications,” *Numerical Algebra, Control and Optimization*, vol. 12, pp. 159–212, 1 2022.

[43] L. E. Ghaoui, M. Oks, and F. Oustry, “Worst-case value-at-risk and robust portfolio optimization: A conic programming approach,” *Operations research*, vol. 51, pp. 543–556, 4 2003.

[44] I. Popescu, “A semidefinite programming approach to optimal-moment bounds for convex classes of distributions,” *Mathematics of Operations Research*, vol. 30, pp. 632–657, 3 2005.

[45] G. Pflug and D. Wozabal, “Ambiguity in portfolio selection,” *Quantitative Finance*, vol. 7, pp. 435–442, 4 2007.

[46] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations research*, vol. 58, pp. 595–612, 3 2010.

[47] Z. Hu and L. J. Hong, "Kullback-leibler divergence constrained distributionally robust optimization," *Available at Optimization Online*, vol. 1, p. 9, 2 2013.

[48] G. A. Hanasusanto and D. Kuhn, "Robust data-driven dynamic programming," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[49] C. Ning and F. You, "Data-driven decision making under uncertainty integrating robust optimization with principal component analysis and kernel smoothing methods," *Computers & Chemical Engineering*, vol. 112, pp. 190–210, 2018.

[50] J. Y.-M. Li, "Closed-form solutions for worst-case law invariant risk measures with application to robust portfolio optimization," *Operations Research*, vol. 66, pp. 1533–1541, 6 2018.

[51] C. Shang and F. You, "Robust optimization in high-dimensional data space with support vector clustering," *IFAC-PapersOnLine*, vol. 51, pp. 19–24, 18 2018.

[52] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, pp. 115–166, 1 2018.

[53] Z. Chen, "Adaptive robust optimization with scenario-wise ambiguity sets," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:53359371.

[54] Z. Chen, D. Kuhn, and W. Wiesemann, "Data-driven chance constrained programs over wasserstein balls," *Operations Research*, 2022.

[55] G. Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[56] L. V. Kantorovich, "On the translocation of masses," vol. 37, 1942, pp. 199–201.

[57] Y. Brenier, "Polar factorization and monotone rearrangement of vector-valued functions," *Communications on pure and applied mathematics*, vol. 44, pp. 375–417, 4 1991.

[58] C. Villani, *Topics in optimal transportation.* American Mathematical Soc., 2021, vol. 58.

[59] A. G. Wilson, "The use of entropy maximising models, in the theory of trip distribution, mode split and route split," *Journal of transport economics and policy*, pp. 108–126, 1969.

[60] S. Erlander, *Optimal spatial interaction and the gravity model.* Springer, 1980, vol. 173.

[61] R. Sinkhorn, "A relationship between arbitrary positive matrices and doubly stochastic matrices," *The annals of mathematical statistics*, vol. 35, pp. 876–879, 2 1964.

[62] R. Sinkhorn and P. Knopp, “Concerning nonnegative matrices and doubly stochastic matrices,” *Pacific Journal of Mathematics*, vol. 21, pp. 343–348, 2 1967.

[63] R. Sinkhorn, “Diagonal equivalence to matrices with prescribed row and column sums,” *The American Mathematical Monthly*, vol. 74, pp. 402–405, 4 1967.

[64] R. Cominetti and J. S. Martin, “Asymptotic analysis of the exponential penalty trajectory in linear programming,” *Mathematical Programming*, vol. 67, pp. 169–187, 1994.

[65] Y. Chen, T. T. Georgiou, and A. Tannenbaum, “Optimal transport for gaussian mixture models,” *IEEE Access*, vol. 7, pp. 6269–6278, 2018.

[66] S. Roweis and Z. Ghahramani, “A unifying review of linear gaussian models,” *Neural computation*, vol. 11, pp. 305–345, 2 1999.

[67] N Kostantinos, “Gaussian mixtures and their applications to signal processing,” *Advanced signal processing handbook: theory and implementation for radar, sonar, and medical imaging real time systems*, pp. 1–3, 2000.

[68] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, “Finite mixture models,” *Annual review of statistics and its application*, vol. 6, pp. 355–378, 2019.

[69] A. Takatsu, “Wasserstein geometry of gaussian measures,” 2011.

[70] H. Heitsch, W. Römisch, and C. Strugarek, “Stability of multistage stochastic programs,” *SIAM Journal on Optimization*, vol. 17, pp. 511–525, 2 2006.

[71] H. Heitsch and W. Römisch, “Scenario tree modeling for multistage stochastic programs,” *Mathematical Programming*, vol. 118, pp. 371–406, 2009.

[72] D. Xu, Z. Chen, and L. Yang, “Scenario tree generation approaches using k-means and lp moment matching methods,” *Journal of Computational and Applied Mathematics*, vol. 236, pp. 4561–4579, 17 2012.

[73] Z. Chen and Z. Yan, “Practical arbitrage-free scenario tree reduction methods and their applications in financial optimization,” *Applied Stochastic Models in Business and Industry*, vol. 34, pp. 175–195, 2 2018.

[74] Z. Li and C. A. Floudas, “Optimal scenario reduction framework based on distance of uncertainty distribution and output performance: I. single reduction via mixed integer linear optimization,” *Computers & Chemical Engineering*, vol. 70, pp. 50–66, 2014.

[75] Z. Li and Z. Li, “Linear programming-based scenario reduction using transportation distance,” *Computers & Chemical Engineering*, vol. 88, pp. 50–58, 2016.

[76] Y. Zhou, L. Shi, and Y. Ni, “An improved scenario reduction technique and its application in dynamic economic dispatch incorporating wind power,” 2019, pp. 3168–3178.

[77] R. Karuppiah, M. Martin, and I. E. Grossmann, “A simple heuristic for reducing the number of scenarios in two-stage stochastic programming,” *Computers & chemical engineering*, vol. 34, pp. 1246–1255, 8 2010.

[78] L. A. A. Meira, G. P. Coelho, A. A. S. Santos, and D. J. Schiozer, “Selection of representative models for decision analysis under uncertainty,” *Computers & Geosciences*, vol. 88, pp. 67–82, 2016.

[79] S. Arpón, T. H. de Mello, and B. Pagnoncelli, “Scenario reduction for stochastic programs with conditional value-at-risk,” *Mathematical Programming*, vol. 170, pp. 327–356, 1 2018.

[80] J. Hu and H. Li, “A new clustering approach for scenario reduction in multi-stochastic variable programming,” *IEEE Transactions on Power Systems*, vol. 34, pp. 3813–3825, 5 2019.

[81] J. Silvente, L. G. Papageorgiou, and V. Dua, “Scenario tree reduction for optimisation under uncertainty using sensitivity analysis,” *Computers & Chemical Engineering*, vol. 125, pp. 449–459, 2019.

[82] Q. Li and D. W. Gao, “Fast scenario reduction for power systems by deep learning,” *arXiv preprint arXiv:1908.11486*, 2019.

[83] S. Medina-Gonzalez, I. Gkioulekas, V. Dua, and L. G. Papageorgiou, “A graph theory approach for scenario aggregation for stochastic optimisation,” *Computers & Chemical Engineering*, vol. 137, p. 106 810, 2020.

[84] G. L. Bounitsis, L. G. Papageorgiou, and V. M. Charitopoulos, “Data-driven scenario generation for two-stage stochastic programming,” *Chemical Engineering Research and Design*, vol. 187, pp. 206–224, 2022.

[85] G. C. Pflug and A. Pichler, “A distance for multistage stochastic optimization models,” *SIAM Journal on Optimization*, vol. 22, pp. 1–23, 1 2012.

[86] J. Lellmann, D. A. Lorenz, C. Schonlieb, and T. Valkonen, “Imaging with kantorovich–rubinstein discrepancy,” *SIAM Journal on Imaging Sciences*, vol. 7, pp. 2833–2859, 4 2014.

[87] M. Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transport,” *Advances in neural information processing systems*, vol. 26, 2013.

[88] G. Peyré, M. Cuturi, *et al.*, “Computational optimal transport: With applications to data science,” *Foundations and Trends in Machine Learning*, vol. 11, pp. 355–607, 5-6 2019.

[89] A. Nemirovski and U. Rothblum, “On complexity of matrix scaling,” *Linear Algebra and its Applications*, vol. 302, pp. 435–460, 1999.

[90] W. E. Deming and F. F. Stephan, “On a least squares adjustment of a sampled frequency table when the expected marginal totals are known,” *The Annals of Mathematical Statistics*, vol. 11, pp. 427–444, 4 1940.

[91] M. Bacharach, “Estimating nonnegative matrices from marginal data,” *International Economic Review*, vol. 6, pp. 294–310, 3 1965.

[92] A. J. Kleywegt, A. Shapiro, and T. H. de Mello, "The sample average approximation method for stochastic discrete optimization," *SIAM Journal on Optimization*, vol. 12, pp. 479–502, 2 2002.

[93] P. Bonami and J. Lee, "Bonmin users' manual," Feb. 2011.

[94] F. Beltrán, W. de Oliveira, and E. C. Finardi, "Application of scenario tree reduction via quadratic process to medium-term hydrothermal scheduling problem," *IEEE Transactions on Power Systems*, vol. 32, pp. 4351–4361, 6 2017.

[95] J. R. Birge and F. Louveaux, *Introduction to stochastic programming.* Springer Science & Business Media, 2011.

[96] J.-P. Watson, D. L. Woodruff, and W. E. Hart, "Pysp: Modeling and solving stochastic programs in python," *Mathematical Programming Computation*, vol. 4, pp. 109–149, 2 2012.

[97] M. Kaut and W Stein, *Evaluation of scenario-generation methods for stochastic programming.* Humboldt-Universität zu Berlin, Mathematisch - Naturwissenschaftliche Fakultät . . ., 2003.

[98] S. Subrahmanyam, J. F. Pekny, and G. V. Reklaitis, "Design of batch chemical plants under market uncertainty," *Industrial & Engineering Chemistry Research*, vol. 33, pp. 2688–2701, 11 1994.

[99] S. W. Wallace and W. T. Ziemba, *Applications of stochastic programming.* SIAM, 2005.

[100] A. Shapiro and A. Nemirovski, "On complexity of stochastic programming problems," *Continuous optimization: Current trends and modern applications*, pp. 111–146, 2005.

[101] A. Ben-Tal and A. Nemirovski, "Robust convex optimization," *Mathematics of operations research*, vol. 23, pp. 769–805, 4 1998.

[102] H. Scarf, "A min max solution of an inventory problem," *Studies in the mathematical theory of inventory and production*, 1958.

[103] H. Rahimian and S. Mehrotra, "Distributionally robust optimization: A review," *arXiv preprint arXiv:1908.05659*, 2019.

[104] W. Wiesemann, D. Kuhn, and M. Sim, "Distributionally robust convex optimization," *Operations Research*, vol. 62, pp. 1358–1376, 6 2014.

[105] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent measures of risk," *Mathematical finance*, vol. 9, pp. 203–228, 3 1999.

[106] A. Ruszczyński and A. Shapiro, "Optimization of convex risk functions," *Mathematics of operations research*, vol. 31, pp. 433–452, 3 2006.

[107] A. R. Hota, A. Cherukuri, and J. Lygeros, "Data-driven chance constrained optimization under wasserstein ambiguity sets," 2019, pp. 1501–1506.

[108] S.-B. Yang and Z. Li, "Kernel distributionally robust chance-constrained process optimization," *Computers & Chemical Engineering*, vol. 165, p. 107 953, 2022.

[109] D. Goldfarb and G. Iyengar, "Robust portfolio selection problems," *Mathematics of operations research*, vol. 28, pp. 1–38, 1 2003.

[110] P. D. Grunwald and A. P. Dawid, "Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory," *arXiv preprint math/0410076*, 2004.

[111] J. Goh and M. Sim, "Distributionally robust optimization and its tractable approximations," *Operations research*, vol. 58, pp. 902–917, 4-part-1 2010.

[112] K. Natarajan and C.-P. Teo, "On reduced semidefinite programs for second order moment bounds with applications," *Mathematical Programming*, vol. 161, pp. 487–518, 2017.

[113] B. P. G. V. Parys, P. J. Goulart, and D. Kuhn, "Generalized gauss inequalities via semidefinite programming," *Mathematical Programming*, vol. 156, pp. 271–302, 2016.

[114] D. Bertsimas and N. Kallus, "From predictive to prescriptive analytics," *Management Science*, vol. 66, pp. 1025–1044, 3 2020.

[115] J.-J. Zhu, W. Jitkrittum, M. Diehl, and B. Schölkopf, "Kernel distributionally robust optimization: Generalized duality theorem and stochastic approximation," 2021, pp. 280–288.

[116] S. S. Abadeh, P. M. M. Esfahani, and D. Kuhn, "Distributionally robust logistic regression," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[117] R. Jiang and Y. Guan, "Risk-averse two-stage stochastic program with distributional ambiguity," *Operations Research*, vol. 66, pp. 1390–1405, 5 2018.

[118] R. Gao and A. Kleywegt, "Distributionally robust stochastic optimization with wasserstein distance," *Mathematics of Operations Research*, 2022.

[119] J. Blanchet, K. Murthy, and F. Zhang, "Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes," *Mathematics of Operations Research*, vol. 47, pp. 1500–1529, 2 2022.

[120] S. Mehrotra and H. Zhang, "Models and algorithms for distributionally robust least squares problems," *Mathematical Programming*, vol. 146, pp. 123–141, 1-2 2014.

[121] J. Y.-M. Li and T. Mao, "A general wasserstein framework for data-driven distributionally robust optimization: Tractability and applications," *arXiv preprint arXiv:2207.09403*, 2022.

[122] H. Liu, J. Qiu, and J. Zhao, "A data-driven scheduling model of virtual power plant using wasserstein distributionally robust optimization," *International Journal of Electrical Power & Energy Systems*, vol. 137, p. 107 801, 2022.

[123] C. Clason, D. A. Lorenz, H. Mahler, and B. Wirth, "Entropic regularization of continuous optimal transport problems," *Journal of Mathematical Analysis and Applications*, vol. 494, p. 124 432, 1 2021.

[124] J.-D. Benamou, “Numerical resolution of an “unbalanced” mass transport problem,” *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 37, pp. 851–868, 5 2003.

[125] L. A. Caffarelli and R. J. McCann, “Free boundaries in optimal transport and monge-ampere obstacle problems,” *Annals of mathematics*, pp. 673–730, 2010.

[126] M. Blondel, V. Seguy, and A. Rolet, “Smooth and sparse optimal transport,” 2018, pp. 880–889.

[127] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, “Scaling algorithms for unbalanced optimal transport problems,” *Mathematics of Computation*, vol. 87, pp. 2563–2609, 314 2018.

[128] V. I. Oliker and L. D. Prussner, “On the numerical solution of the equation and its discretizations, i,” *Numerische Mathematik*, vol. 54, pp. 271–293, 3 1989.

[129] Q. Mérigot, “A multiscale approach to optimal transport,” *Computer Graphics Forum*, vol. 30, pp. 1583–1592, 5 2011.

[130] B. Lévy, “A numerical algorithm for l$_{\{}2\}semi-discreteoptimaltransportin3d$,” *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, vol. 49, pp. 1693–1715, 6 2015.

[131] B. Pass, “On the local structure of optimal measures in the multi-marginal optimal transportation problem,” *Calculus of Variations and Partial Differential Equations*, vol. 43, pp. 529–536, 3-4 2012.

[132] B. Pass, “Multi-marginal optimal transport: Theory and applications,” *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, vol. 49, pp. 1771–1790, 6 2015.

[133] L. Nenna, “Numerical methods for multi-marginal optimal transportation,” 2016.

[134] I. Haasler, R. Singh, Q. Zhang, J. Karlsson, and Y. Chen, “Multi-marginal optimal transport and probabilistic graphical models,” *IEEE Transactions on Information Theory*, vol. 67, pp. 4647–4668, 7 2021.

[135] B. Aragam, C. Dan, P. Ravikumar, and E. P. Xing, “Identifiability of nonparametric mixture models and bayes optimal clustering,” *arXiv preprint arXiv:1802.04397*, 2018.

[136] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the royal statistical society: series B (methodological)*, vol. 39, pp. 1–22, 1 1977.

[137] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning.* Springer, 2006.

[138] L. You, H. Ma, T. K. Saha, and G. Liu, “Gaussian mixture model based distributionally robust optimal power flow with cvar constraints,” *arXiv preprint arXiv:2110.13336*, 2021.

[139] M.-C. Yue, D. Kuhn, and W. Wiesemann, “On linear optimization over wasserstein balls,” *Mathematical Programming*, vol. 195, pp. 1107–1122, 1 2022.

[140] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski, *Robust optimization.* Princeton university press, 2009, vol. 28.

[141] J. M. Keynes, *A treatise on probability.* Courier Corporation, 2013.

[142] H. E. Scarf, K. J. Arrow, and S Karlin, *A min-max solution of an inventory problem.* Rand Corporation Santa Monica, 1957.

[143] S. Mehrotra and D. Papp, “A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization,” *SIAM Journal on Optimization*, vol. 24, pp. 1670–1697, 4 2014.

[144] M. Bansal, K.-L. Huang, and S. Mehrotra, “Decomposition algorithms for two-stage distributionally robust mixed binary programs,” *SIAM Journal on Optimization*, vol. 28, pp. 2360–2383, 3 2018.

[145] D. Bertsimas, X. V. Doan, K. Natarajan, and C.-P. Teo, “Models for minimax stochastic linear optimization problems with risk aversion,” *Mathematics of Operations Research*, vol. 35, pp. 580–602, 3 2010.

[146] A. Ben-Tal, D. D. Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen, “Robust solutions of optimization problems affected by uncertain probabilities,” *Management Science*, vol. 59, pp. 341–357, 2 2013.

[147] N. Noyan, G. Rudolf, and M. Lejeune, “Distributionally robust optimization with decision-dependent ambiguity set,” *Optimization Online*, 2018.

[148] C. Villani *et al.*, *Optimal transport: old and new.* Springer, 2009, vol. 338.

[149] J. Blanchet and K. Murthy, “Quantifying distributional model risk via optimal transport,” *Mathematics of Operations Research*, vol. 44, pp. 565–600, 2 2019.

[150] R. Gao and A. Kleywegt, “Distributionally robust stochastic optimization with wasserstein distance,” *Mathematics of Operations Research*, vol. 48, pp. 603–655, 2 2023.

[151] G. A. Hanasusanto and D. Kuhn, “Conic programming reformulations of two-stage distributionally robust linear programs over wasserstein balls,” *Operations Research*, vol. 66, pp. 849–869, 3 2018.

[152] F. Luo and S. Mehrotra, “Decomposition algorithm for distributionally robust optimization using wasserstein metric with an application to a class of regression models,” *European Journal of Operational Research*, vol. 278, pp. 20–35, 1 2019.

[153] R. Jiang and Y. Guan, “Data-driven chance constrained stochastic program,” *Mathematical Programming*, vol. 158, pp. 291–327, 1-2 2016.

[154] W. Xie and S. Ahmed, “On deterministic reformulations of distributionally robust joint chance constrained optimization problems,” *SIAM Journal on Optimization*, vol. 28, pp. 1151–1182, 2 2018.

[155] R. Ji and M. A. Lejeune, “Data-driven distributionally robust chance-constrained optimization with wasserstein metric,” *Journal of Global Optimization*, vol. 79, pp. 779–811, 4 2021.

[156] J. Delon and A. Desolneux, “A wasserstein-type distance in the space of gaussian mixture models,” *SIAM Journal on Imaging Sciences*, vol. 13, pp. 936–970, 2 2020.

[157] G. Dusson, V. Ehrlacher, and N. Nouaime, “A wasserstein-type metric for generic mixture models, including location-scatter and group invariant measures,” *arXiv preprint arXiv:2301.07963*, 2023.

[158] A. Prékopa, “Probabilistic programming,” *Handbooks in operations research and management science*, vol. 10, pp. 267–351, 2003.

[159] M. Y. An, “Log-concave probability distributions: Theory and statistical testing,” *Duke University Dept of Economics Working Paper*, 95-03 1997.

[160] B. K. Pagnoncelli, S. Ahmed, and A. Shapiro, “Sample average approximation method for chance constrained programming: Theory and applications,” *Journal of optimization theory and applications*, vol. 142, pp. 399–416, 2 2009.

[161] S.-B. Yang and Z. Li, “Distributionally robust chance-constrained optimization with sinkhorn ambiguity set,” *AIChE Journal*, vol. 69, e18177, 10 2023.

[162] R. N. Sauer, A. R. Colville, and C. W. Burwick, “Computer points way to more profits,” *Hydrocarbon Processing*, vol. 84, 2 1964.

[163] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, “A review of process fault detection and diagnosis: Part i: Quantitative model-based methods,” *Computers & chemical engineering*, vol. 27, pp. 293–311, 3 2003.

[164] S. Aminikhanghahi and D. J. Cook, “A survey of methods for time series change point detection,” *Knowledge and information systems*, vol. 51, pp. 339–367, 2 2017.

[165] F. Desobry, M. Davy, and C. Doncarli, “An online kernel change detection algorithm,” *IEEE Transactions on Signal Processing*, vol. 53, pp. 2961–2974, 8 2005.

[166] M. Han, L. T. Vinh, Y.-K. Lee, and S. Lee, “Comprehensive context recognizer based on multimodal sensors in a smartphone,” *Sensors*, vol. 12, pp. 12 588–12 605, 9 2012.

[167] Y. Kawahara and M. Sugiyama, “Change-point detection in time-series data by direct density-ratio estimation,” 2009, pp. 389–400.

[168] L. I. Kuncheva, “Change detection in streaming multivariate data using likelihood detectors,” *IEEE transactions on knowledge and data engineering*, vol. 25, pp. 1175–1180, 5 2011.

[169] V. Venkatasubramanian, R. Rengaswamy, and S. N. Kavuri, “A review of process fault detection and diagnosis: Part ii: Qualitative models and search strategies,” *Computers & chemical engineering*, vol. 27, pp. 313–326, 3 2003.

[170] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, and K. Yin, "A review of process fault detection and diagnosis: Part iii: Process history based methods," *Computers & chemical engineering*, vol. 27, pp. 327–346, 3 2003.

[171] D. Garcia-Alvarez, M. J. Fuente, P. Vega, and G. Sainz, "Fault detection and diagnosis using multivariate statistical techniques in a wastewater treatment plant.," *IFAC Proceedings Volumes*, vol. 42, pp. 952–957, 11 2009.

[172] J Mina and C Verde, "Fault detection using dynamic principal component analysis by average estimation," 2005, pp. 374–377.

[173] B. R. Bakshi, "Multiscale pca with application to multivariate statistical process monitoring," *AIChE journal*, vol. 44, pp. 1596–1610, 7 1998.

[174] R. T. Samuel and Y. Cao, "Fault detection in a multivariate process based on kernel pca and kernel density estimation," 2014, pp. 146–151.

[175] Y. Zhang and Q. Jia, "Complex process monitoring using kuca with application to treatment of waste liquor," *IEEE Transactions on Control Systems Technology*, vol. 26, pp. 427–438, 2 2017.

[176] F. Harrou, Y. Sun, and M. Madakyaru, "Kullback-leibler distance-based enhanced detection of incipient anomalies," *Journal of Loss Prevention in the Process Industries*, vol. 44, pp. 73–87, 2016.

[177] F. Harrou, M. Madakyaru, and Y. Sun, "Improved nonlinear fault detection strategy based on the hellinger distance metric: Plug flow reactor monitoring," *Energy and Buildings*, vol. 143, pp. 149–161, 2017.

[178] M. Gelbrich, "On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces," *Mathematische Nachrichten*, vol. 147, pp. 185–203, 1 1990.

[179] T. Rippl, A. Munk, and A. Sturm, "Limit laws of the empirical wasserstein distance: Gaussian distributions," *Journal of Multivariate Analysis*, vol. 151, pp. 90–109, 2016.

[180] B. M. S. Arifin, Z. Li, and S. L. Shah, "Change point detection using the kantorovich distance algorithm," *IFAC-PapersOnLine*, vol. 51, pp. 708–713, 18 2018.

[181] F. Harrou, M. Madakyaru, Y. Sun, and S. Kammammettu, "Enhanced dynamic data-driven fault detection approach: Application to a two-tank heater system," 2017, pp. 1–6.

[182] J. E. Jackson and G. S. Mudholkar, "Control procedures for residuals associated with principal component analysis," *Technometrics*, vol. 21, pp. 341–349, 3 1979.

[183] T. Villegas, M. J. Fuente, and M. Rodr ́, "Principal component analysis for fault detection and diagnosis. experience with a pilot plant," *Advances in Computational Intelligence, Man-Machine Systems and Cybernetics*, pp. 147–152, 2010.

[184] J. Harmouche, “Statistical incipient fault detection and diagnosis with kullback-leibler divergence: From theory to applications,” 2014.

[185] N. F. Thornhill, S. C. Patwardhan, and S. L. Shah, “A continuous stirred tank heater simulation model with applications,” *Journal of process control*, vol. 18, pp. 347–360, 3-4 2008.

[186] J. J. Downs and E. F. Vogel, “A plant-wide industrial process control problem,” *Computers & chemical engineering*, vol. 17, pp. 245–255, 3 1993.

[187] C. Shang, S. X. Ding, and H. Ye, “Distributionally robust fault detection design and assessment for dynamical systems,” *Automatica*, vol. 125, p. 109 434, 2021.

[188] Y. Wan, Y. Ma, and M. Zhong, “Distributionally robust trade-off design of parity relation based fault detection systems,” *International Journal of Robust and Nonlinear Control*, vol. 31, no. 18, pp. 9149–9174, 2021.

[189] J. Yu and S. J. Qin, “Multimode process monitoring with bayesian inference-based finite gaussian mixture models,” *AIChE Journal*, vol. 54, no. 7, pp. 1811–1829, 2008.

[190] S. W. Choi, J. H. Park, and I.-B. Lee, “Process monitoring using a gaussian mixture model via principal component analysis and discriminant analysis,” *Computers & chemical engineering*, vol. 28, no. 8, pp. 1377–1387, 2004.

[191] J. Zhang, D. Zhou, and M. Chen, “Monitoring multimode processes: A modified pca algorithm with continual learning ability,” *Journal of Process Control*, vol. 103, pp. 76–86, 2021.

[192] K. Zhang, *Performance assessment for process monitoring and fault detection methods*. Springer, 2016.

[193] Y. Wang and C. A. Shoemaker, “A general stochastic algorithmic framework for minimizing expensive black box objective functions based on surrogate models and sensitivity analysis,” *arXiv preprint arXiv:1410.6271*, 2014.

[194] R. G. Regis and C. A. Shoemaker, “A stochastic radial basis function method for the global optimization of expensive functions,” *INFORMS Journal on Computing*, vol. 19, no. 4, pp. 497–509, 2007.

Appendix A: Wasserstein ambiguity set-based DRCCP

In this section, we give the derivation of a tractable formulation for distributionally robust chance-constrained programming problem (DRCCP) using the 1-Wasserstein distance-based ambiguity set. The derivation procedure is similar to that of Yang and Li (2022), and we include an overview of the same to ensure completeness of this work. We start from the worst-case expectation problem in the constraint of DRCCP Model 4.16.

$$\max_{\mathbb{P}\in\mathcal{P}} \ \mathbb{E}_{\mathbb{P}}\Big(\max_{1\leq i\leq m} \ g_i(x,\xi)-\eta\Big)^+$$

Under the 1-Wasserstein metric-based ambiguity set of radius ϵ_W, computed through the discrete optimal transport problem (Model 4.8), wherein the expectation operator is empirically estimated over a candidate distribution supported on H points, the model may be further formulated as,

$$\max_{\pi_{j,h}\geq 0,\rho_h\geq 0} \quad \sum_{h=1}^{H}\rho_h\Big(\max_{1\leq i\leq m} \ g_i(x,\xi)-\eta\Big)^+ \tag{A.1a}$$

$$\text{s.t.} \quad \min_{\pi_{j,h},\rho_h} \quad \sum_{j=1}^{N}\sum_{h=1}^{H}||\xi_h-\xi_j||\pi_{j,h}\leq\epsilon_W \tag{A.1b}$$

$$\sum_{h=1}^{H}\pi_{j,h}=\frac{1}{N}, \quad 1\leq j\leq N \tag{A.1c}$$

$$\sum_{j=1}^{N}\pi_{j,h}=\rho_h, \quad 1\leq h\leq H \tag{A.1d}$$

Since this worst-case expectation problem forms the inner maximization problem in the overall DRCCP minimization problem, we can denote $\mathcal{L}(x,\xi) \coloneqq \Big(\max_{1\leq i\leq m} \ g_i(x,\xi)-$

$\eta)^{+}$. Model A.1 may be further written in terms of the transportation variable π only as follows, while dropping the "min" operator,

$$\max_{\pi_{j,h}\geq 0,\rho_h\geq 0} \quad \sum_{j=1}^{N}\sum_{h=1}^{H} \rho_h \mathcal{L}(x,\xi_h) \tag{A.2a}$$

$$\text{s.t.} \quad \sum_{j=1}^{N}\sum_{h=1}^{H} ||\xi_h - \xi_j||\pi_{j,h} \leq \epsilon_W \tag{A.2b}$$

$$\sum_{h=1}^{H} \pi_{j,h} = \frac{1}{N}, \quad 1 \leq j \leq N \tag{A.2c}$$

The dual problem for this model may be obtained by introducing the dual variables κ and $z_j, \quad 1 \leq j \leq N$ as follows,

$$\min_{\kappa\geq 0,z_j} \quad \kappa\epsilon_W + \frac{1}{N}\sum_{j=1}^{N} z_j \tag{A.3a}$$

$$\text{s.t.} \quad z_j \geq \mathcal{L}(x,\xi_h) - \kappa||\xi_h - \xi_j||, \quad 1 \leq j \leq N, 1 \leq h \leq H \tag{A.3b}$$

In practice, the candidate distributions themselves as well as their supports are unknown to the user, and therefore, an infinite continuous support $\xi \in \Xi$ is established as follows,

$$\min_{\kappa\geq 0,z_j} \quad \kappa\epsilon_W + \frac{1}{N}\sum_{j=1}^{N} z_j \tag{A.4a}$$

$$\text{s.t.} \quad z_j \geq \max_{\xi\in\Xi} \quad \mathcal{L}(x,\xi) - \kappa||\xi - \xi_j||, \quad 1 \leq j \leq N \tag{A.4b}$$

This model may be further reformulated using the dual norm as $-\kappa||\xi - \xi_j^0|| = -\min_{||V_j||_*\leq\kappa} \quad V_j^T(\xi - \xi_j)$ and dropping the "min" operator,

$$\min_{\kappa,z_j} \quad \kappa\epsilon_W + \frac{1}{N}\sum_{j=1}^{N} z_j \tag{A.5a}$$

$$\text{s.t.} \quad z_j \geq \max_{\xi\in\Xi} \quad \mathcal{L}(x,\xi) - V_j^T(\xi - \xi_j), \quad 1 \leq j \leq N \tag{A.5b}$$

$$||V_j||_* \leq \kappa, \quad 1 \leq j \leq N \tag{A.5c}$$

For a piece-wise linear loss function that may be written as $\mathcal{L}(x,\xi) := \max\limits_{1\leq i\leq m} \quad a_i\xi + b_i$, the model may be reformulated as,

$$\min_{\kappa,z_j} \quad \kappa\epsilon_W + \frac{1}{N}\sum_{j=1}^{N} z_j \tag{A.6a}$$

$$\text{s.t.}\quad z_j \geq \max_{\xi \in \Xi} \left(a_i \xi - V_j^T \xi\right) + b_i + V_j^T \xi_j, \quad 1 \leq j \leq N, 1 \leq i \leq m \tag{A.6b}$$

$$\|V_j\|_* \leq \kappa, \quad 1 \leq j \leq N \tag{A.6c}$$

The inner maximization problem in A.6b may be written as $\max_{\xi \in \Xi} \quad (a_i - V_j)^T \xi$ for each i, j. If Ξ is the real space, then the inner maximization problem is an unbounded linear programming problem unless $a_i = V_j, \quad \forall i, j$. Therefore, the model may be reformulated as,

$$\min_{\kappa, z_j} \quad \kappa \epsilon_W + \frac{1}{N} \sum_{j=1}^{N} z_j \tag{A.7a}$$

$$\text{s.t.}\quad z_j \geq b_i + a_i^T \xi_j^0, \quad 1 \leq j \leq N, 1 \leq i \leq m \tag{A.7b}$$

$$\|a_i\|_* \leq \kappa, \quad 1 \leq i \leq m \tag{A.7c}$$

If the constraint function in the joint chance constraint is affine: $g_i(x, \xi) := h_i(x)\xi + \overline{h_i}(x) \quad \forall i$, then $\mathcal{L}(x, \xi) := \left(\max_{1 \leq i \leq m} \; g_i(x, \xi) - \eta\right)^+ := \left(\max_{1 \leq i \leq m} \; h_i(x)\xi + \overline{h_i}(x) - \eta\right)^+$, Model A.7 may be reformulated with $a_i = h_i(x)$, $b_i = \overline{h_i}(x) - \eta, \quad 1 \leq i \leq m$ and $a_{m+1} = 0, b_{m+1} = 0$. Substitute it into Model 4.16 and drop the "min" operator, we have

$$\min_{x \in X, \eta, \kappa, z_j} \quad f(x) \tag{A.8a}$$

$$\text{s.t.}\quad \eta + \frac{1}{\delta}\left[\kappa \epsilon_W + \frac{1}{N} \sum_{j=1}^{N} z_j\right] \leq 0 \tag{A.8b}$$

$$z_j \geq \overline{h_i}(x) - \eta + h_i(x)^T \xi_j^0, \quad 1 \leq j \leq N, 1 \leq i \leq m \tag{A.8c}$$

$$z_j \geq 0, \quad 1 \leq j \leq N \tag{A.8d}$$

$$\|h_i(x)\|_* \leq \kappa, \quad 1 \leq i \leq m \tag{A.8e}$$