REPRESENTATIONS OF UNCERTAINTY
FOR
TRACTABLE ESTIMATION AND CONTROL

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF AERONAUTICS AND ASTRONAUTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Alexandros E. Tzikas
July 2026

Alexandros E. Tzikas

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Mykel J. Kochenderfer) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Marco Pavone)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Mac Schwager)

Approved for the Stanford University Committee on Graduate Studies

Abstract

This thesis studies three representations of uncertainty that are useful for tractable estimation and control. Modern decision making problems are data-driven, high-dimensional, distribution-aware, and sequential. Although such problems are complex, we identify a unifying principle: rather than modeling or optimizing at full fidelity, one can appropriately restrict or summarize the problems, retaining the information most relevant to the task, and obtain good performance. This is done by carefully selecting an adequate representation of the underlying uncertainty. We study three representations of uncertainty in this context. First, we consider a structural representation in the form of a factor model and show how one can refine this tractable model for uncertainty in high dimensions. We apply our method on financial returns and show how it can be used to design better portfolios. Second, we study a geometric representation in the form of a computationally efficient family of distances between probability distributions. We call this a geometric representation because it relies on averaging discrepancies over one-dimensional projections. We present applications in distribution-aware control by designing simple gradient-based controllers that minimize this distance. Third, we study an informational representation that allows updates by conditioning on newly revealed information. We show how to leverage this representation for optimal resource allocation under stochastic demands using a shrinking-horizon optimization scheme. Together, these representations provide a unified framework for learning uncertainty models, enriching control objectives, and adapting decisions over time.

Acknowledgments

Pursuing doctoral studies at Stanford has been one of the best decisions of my life. Looking back, I am now a different person than the one who started this journey. This is thanks to the people I have met along the way: the ones who have supported me, and, when needed, challenged me.

I am grateful to have had the opportunity to work with Prof. Mykel J. Kochenderfer, who has been an incredible mentor and advisor. Mykel, you have always fostered my ambitions and goals. I cannot think of a better advisor.

I also want to thank Prof. Stephen P. Boyd, Prof. Emmanuel J. Candès, and Prof. Trevor Hastie. Without your guidance I would be a much lesser researcher. Stephen, thank you for teaching me the value of simplicity and abstraction. Emmanuel, thank you for teaching me the value of mathematical elegance and first principles. Trevor, thank you for teaching me the value of intuition. Our conversations have been invaluable and will continue to fuel my desire for learning and research. I hope to walk in your footsteps.

I am also indebted to Prof. Marco Pavone and Prof. Mac Schwager for their support and feedback. I have learned a lot from you and you have been most influential in my research interests and journey.

To all my collaborators, past and present, thank you for the opportunity to work with you and for the many things you have taught me. I am grateful for the opportunity to have worked with such talented and kind people.

Finally, to my family and friends, thank you for your love and support. I am grateful to have you in my life and I hope to have made you proud.

Contents

List of Tables

List of Figures

List of Algorithms

1 Introduction

One of the most fundamental problems in artificial intelligence is the control of an agent under uncertainty. This thesis focuses on developing algorithms for tractable (i) modeling of uncertainty and (ii) control under uncertainty that are suitable for modern decision making, namely data-driven, high-dimensional, distribution-aware, and sequential decision making problems. We study different representations of uncertainty that extract the relevant information from the data and allow us to design tractable algorithms for estimation and control. This chapter discusses the challenges of modern decision making under uncertainty, presents a unifying problem that captures these challenges, summarizes our contributions, and outlines the rest of the thesis.

1.1 Modern decision making under uncertainty

Many important applications involve decision making under uncertainty. Notable examples include financial portfolio construction, autonomous driving, aircraft collision avoidance, and robotic exploration. In these settings, an *agent* must reason about the consequences of their actions in the face of uncertainty about the state of the world and the effects of their actions on the future state of the world. We define an *agent* to be any entity that

- acts based on its information about the state of the world,
- seeks to optimize an objective, and
- its decisions have an effect on the future state of the world.

In this thesis, we use the terms *control*, *planning*, and *decision making* interchangeably to refer to the process of selecting actions by an agent in order to optimize an objective in the face of uncertainty.

Uncertainty enters the decision making problem through two qualitatively different sources. First, there is inherent uncertainty in the evolution of the state of the world. For example, financial returns are inherently unpredictable and their dynamics are not deterministic. This type of uncertainty is often called *aleatoric* uncertainty. Aleatoric uncertainty cannot be mitigated, but the agent can learn to make good decisions in the face of it. There is also uncertainty in the agent's knowledge about the state of the world. For example, an autonomous vehicle may have noisy sensors that provide imperfect information about its surroundings. This type of uncertainty is often called *epistemic* uncertainty. Epistemic uncertainty can be reduced through statistical learning and exploration, but it usually cannot be completely eliminated.

The history of decision making under uncertainty is rich and spans multiple disciplines, including economics, operations research, control theory, and artificial intelligence. However, recent advancements in computing capabilities and data collection have led to a paradigm shift. Modern decision making problems are

- **Data-driven**. There is an abundance of available data to assist the learning and decision making of the agent. For example, in financial portfolio construction, there is a wealth of historical data on asset returns. In autonomous driving, there is a wealth of sensor data from vehicles and infrastructure. This data can be used to learn models of the world and to make informed decisions. However, leveraging the data effectively is a significant challenge. Typically, the inherent uncertainty of the problem allows for spurious correlations to appear in the data that can lead to over-fitting. This is especially true when dealing with high-dimensional data. Uncovering structure from large amounts of data is a key challenge in modern decision making under uncertainty.

- **High-dimensional**. The dimensionality of the state of the world and the decision variable can be very large. For example, in financial portfolio construction,

the state of the world may include the returns of thousands of assets, and the decision variable may be the allocation of capital across those assets. In autonomous driving, the state of the world may include the positions and velocities of all nearby vehicles and pedestrians, and the decision variable may be the trajectory of the vehicle. High-dimensional problems are challenging because they suffer from the curse of dimensionality, which makes it difficult to compute optimal decisions and to learn accurate models of the world as the dimensionality increases.

- **Distribution-aware**. The objective the agent seeks to optimize is often as rich as the underlying uncertainty model, going beyond lower-order statistical moments and relying on the full distribution of outcomes. Such objectives further complicate the planning task, as they require reasoning about the entire distribution of outcomes rather than just a few summary statistics.

- **Sequential**. The agent must make a sequence of decisions over time, and the consequences of those decisions can be long-term. For example, in financial portfolio construction, the agent must make a sequence of investment decisions; one at each rebalance date. There can even be nonstationarity in the evolution of the state of the world. This requires learning a dynamic model of the world. In addition, information is usually revealed in a sequential manner. This sequential nature of decision making under uncertainty requires algorithms that effectively incorporate information over time in the planning process.

To address modern control, we require (i) uncertainty models that are statistically consistent with data and computationally tractable (to avoid over-fitting) and (ii) control algorithms that can efficiently handle high-dimensional problems and expressive objectives, leverage data effectively, and exploit the sequential nature of the problems.

1.2 *A unifying problem: financial portfolio construction*

We summarize the key challenges in modern decision making under uncertainty with a unifying problem: financial portfolio construction. In this problem, an investor seeks to allocate capital across a set of assets in order to maximize the expected return of their portfolio while controlling for risk. The investor must make these decisions in the face of uncertainty about the returns of the assets and the dynamics of the market.

The typical formulation of the portfolio optimization problem is

$$\begin{aligned} \max_{w_t} \quad & \mathbb{E}[r_t]^\top w_t \\ \text{s.t.} \quad & w_t^\top \Sigma_t w_t \leq \sigma^2. \end{aligned} \tag{1.1}$$

where we ignore any additional constraints for simplicity. Those would account for any additional requirements imposed by the investor. The decision variable is the portfolio allocation vector w_t, which represents the fraction of capital allocated to each asset at time t. The objective is to maximize the expected return of the portfolio, which is given by $\mathbb{E}[r_t]^\top w_t$, where $\mathbb{E}[r_t]$ is the estimated expected return of the assets at time t. The constraint is a risk constraint, which limits the variance of the portfolio returns to be less than or equal to σ^2, where Σ_t is the covariance matrix of asset returns estimated at time t.

Both $\mathbb{E}[r_t]$ and Σ_t are latent population quantities: in practice, one observes only a single realization of returns at time t, not the distribution from which those returns are drawn. Nevertheless, the investor can leverage historical data to learn a model of the uncertainty, which can then be used to make informed decisions that are robust to uncertainty.

A typical number of assets in a portfolio is on the order of hundreds to thousands, which makes the problem high-dimensional. The investor typically has access to a large amount of historical data on asset returns, which can be used to learn the model of uncertainty. The investor may also have a more expressive objective that goes beyond maximizing expected return and controlling for variance. For example,

the investor may seek to maximize the probability of achieving a certain return threshold, to minimize the expected shortfall of the portfolio returns, or to steer the distribution of portfolio returns towards a desired target distribution. Finally, the investor can adjust the portfolio allocation over time, as new information about asset returns is revealed and as the market dynamics evolve. The investor would need to re-solve the portfolio optimization problem for this purpose.

1.3 Contributions

This thesis focuses on developing algorithms for tractable (i) modeling of uncertainty and (ii) control under uncertainty that can handle data-driven, high-dimensional, distribution-aware, and sequential decision making problems.

Although many researchers are currently working on these problems and solutions have been proposed for specific instances of modern decision making, in this thesis we aim for a unified treatment. We recognize that selecting an appropriate representation of the uncertainty is key for good performance in any decision making problem.

We study three different representations of uncertainty. Each representation is designed to address a specific aspect of the unifying problem (1.1), and together they provide a comprehensive, yet not exhaustive, framework for modeling and control under uncertainty for modern applications.

1.3.1 Structural representation

The first representation we present is *structural*. This representation is designed to capture the structure of the uncertainty in the problem. Specifically, we focus on the problem of modeling the covariance matrix of a high-dimensional random vector. The covariance matrix is a key component of the uncertainty model, and it is often difficult to estimate accurately in high dimensions. The structural representation for the covariance model we study is the family of low-rank-plus-diagonal matrices.

Low-rank-plus-diagonal models are also known as factor models in the literature

of applied statistics. They are a powerful tool for modeling high-dimensional covariance matrices, as they can capture the underlying structure of the data while also being computationally tractable. Learning statistical factor models is a fundamental problem both in finance and state estimation more broadly. These models are interpretable because they expose shared drivers (also known as factors) behind many variables. The small number of learned parameters reduces estimation variance and over-fitting. Low-rank-plus-diagonal models also allow for fast downstream optimization and inference tasks.

Our contribution is a principled method for extending a provided low-rank-plus-diagonal model for the covariance. This extension is based on the idea of adding a few additional statistical factors to the model in order to capture additional structure in the data. Extending a provided model can be important because

- the provided model may be updated too infrequently to reflect transient factors or regime shifts in the data and
- the provided model may be missing (statistical) factors.

However, a learned factor model only captures second-order uncertainty information. Many control problems do not only involve moments of the distribution, but depend on the full distribution of outcomes. It is therefore important to design scalable methods for measuring the similarity between distributions. This motivates the second representation we present, which we call a geometric representation, and captures the mismatch between two distributions.

1.3.2 Geometric representation

Many distances and divergences between probability distributions have been proposed in the literature, including the Kullback-Leibler divergence, the Wasserstein distance, and the total variation distance. These distances tend to not scale well to high dimensions. Nevertheless, quantifying the mismatch between two distributions is important for control under uncertainty, as it allows us to design objectives that are based on the distribution of outcomes rather than just a few summary statistics.

Our contribution is a computationally efficient, interpretable, and suitable for control applications family of distances between probability distributions. This family of distances is based on the idea of projecting the distributions onto a one-dimensional space and then computing the distance between the projected distributions. This is much easier than directly computing a distance between high-dimensional distributions. Our distance is then the average distance between the projected one-dimensional distributions, averaged over a set of projection directions. We therefore call it a *geometric* representation of uncertainty, as it projects the distributions onto one-dimensional spaces.

We further demonstrate how this family of distances can be used to control the distribution of outcomes in a tractable manner using simple gradient-based controllers.

Although the geometric representation enables distribution-aware control, it does not, on its own, provide a principled way to update decisions as new information arrives. Incorporating newly revealed information offers an additional mechanism for improving control performance. This motivates the third representation of uncertainty we present.

1.3.3 Informational representation

We define an *informational* representation of uncertainty as one that allows us to easily incorporate newly revealed information. We study a method for control that leverages new information by updating the informational uncertainty model before new decisions are made. This method updates the statistical moments of uncertainty by conditioning on the provided information. Assuming $\mathcal{F}_t$ is the information available at time t, we can equivalently view the method as a projection of the original model onto the family of models that are $\mathcal{F}_t$-measurable. The method is computationally efficient and can be used in real-time applications.

Our contribution is as follows. We consider the problem of resource allocation under stochastic demands. We propose an adequate informational representation of uncertainty and combine it with a shrinking horizon control scheme to design a

tractable algorithm to make allocations over time.

1.3.4 *Summary of contributions*

In the context of the unifying problem of financial portfolio construction, the three representations are summarized in Figure 1.1. The structural representation improves the estimate Σ_t. The geometric representation allows for a more expressive objective that goes beyond maximizing expected return. The informational representation allows for easily incorporating newly revealed information, e.g., computing and using $\mathbb{E}[r_t \mid \mathcal{F}_t]$ instead of $\mathbb{E}[r_t]$, into the decision making process. A high-level summary of our three contributions is given in Figure 1.2.

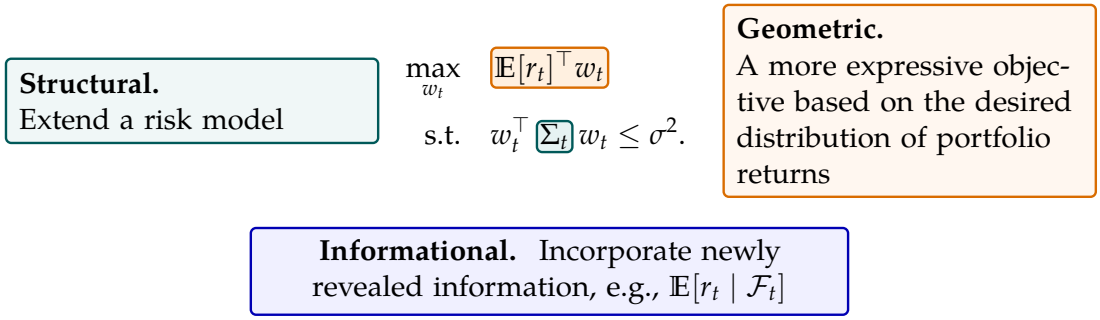


Figure 1.1: A unifying portfolio optimization problem that highlights key aspects of modern decision-making under uncertainty: *structural* modeling of risk (Σ_t), *geometric* (distribution-aware) objectives, and *informational* adaptation as new data arrive.

1.4 *Outline*

The rest of the thesis is organized as follows. In Chapter 2, we present the structural representation of uncertainty and our contribution to learning factor models for covariance estimation. In Chapter 3, we present our geometric representation of uncertainty and its application to distribution-aware control. In Chapter 4, we present the informational representation of uncertainty and our contribution to

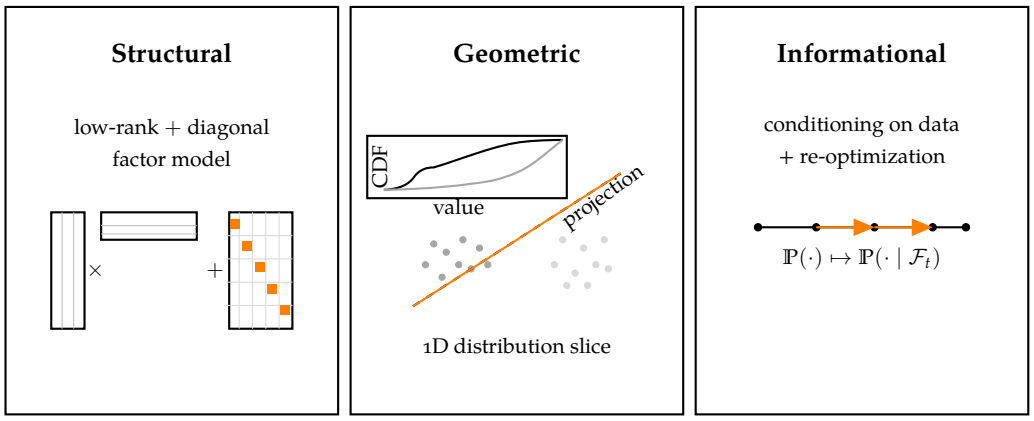


Figure 1.2: The three representations of uncertainty studied in this thesis. *Structural* representation enforces a low-rank-plus-diagonal factor structure for tractable risk modeling, *geometric* representation reduces multivariate distributions to collections of one-dimensional slices for distribution-aware objectives, and *informational* representation conditions on newly revealed data and re-optimizes to adapt decisions over time.

resource allocation under uncertainty. Finally, in Chapter 5, we conclude with a summary of our contributions and directions for future work.

2 *Structural representation*

In this chapter we focus on improving a provided covariance model in low-rank-plus-diagonal form. This specific structure of models for the covariance justifies the name of the chapter. We introduce an algorithm that allows us to extend and refine the provided low-rank-plus-diagonal covariance model by adding new statistical factors. We present the methodology in the context of extending a model for the covariance of financial returns. Nevertheless, the method is general and can be applied to any problem where a low-rank-plus-diagonal model is used.

Estimating the covariance of asset returns, i.e., the risk model, is a key component of financial portfolio construction and evaluation. An accurate risk model allows investors to build risk-aware portfolios and assess the risk level of already constructed ones. Most risk modeling approaches produce a low-rank-plus-diagonal model that decomposes the asset variability into two components: the first attributed to a small number of factors that are common among the assets and the second attributed to the idiosyncratic behavior of each asset. Low-rank-plus-diagonal models are both interpretable and computationally efficient. Most investors rely on third-party providers for their risk models. Although these models are typically of high quality, they may fail to capture important information, e.g., changing market regimes and transient factors. To overcome these limitations, we propose a method based on maximum likelihood estimation to extend a provided low-rank-plus-diagonal risk model by refining the given model and adding new statistical factors. Our method easily extends to the case of missing data.

2.1 Introduction

The covariance of asset returns is a central quantity in several areas of finance, including Markowitz portfolio construction [1], [2], risk management [3], and asset pricing [4]. The risk of a portfolio (or investment strategy) is defined as the variance of its return, where the portfolio return is a linear function of the asset returns vector [5]. Therefore the portfolio risk is a linear function of the asset return covariance. An accurate, as judged by its statistical fit, asset return covariance allows investors to build portfolios with a desired level of risk and also evaluate the risk of existing portfolios. An inaccurate risk model can over-estimate or under-estimate portfolio risk. As a result, during portfolio construction, it may be overly restrictive, causing the investor to miss out on potentially good investments, or fail to account for certain risk sources, potentially exposing the investment strategy to unexpected risks. Risk modeling is therefore a challenging problem in any investment process and requires careful consideration.

A major challenge is that only information available at the current time can be used to estimate the risk model, which should be predictive of the assets' behavior in the future. Prediction is particularly difficult because of changing market conditions [6], extreme events (e.g., COVID-19), and fat tails in the distribution of returns [7]. In addition, asset returns can fluctuate significantly even on a daily basis, making it difficult to separate meaningful structure from randomness [5]. The high dimensionality of the data, i.e., the large number of assets (possibly thousands), is another challenge [8]. For example, the sample covariance based on the observed data is singular when the number of assets is larger than the number of sample return vectors [7]. This is problematic because likelihood calculations involve the inverse of the covariance, i.e., the precision matrix, and the rank deficiency implies that there are directions in asset space with zero variance. The latter can cause problems in portfolio construction. Because of these challenges, various companies, for example MSCI Barra [9], [10], offer risk models as products and even a Nobel Memorial Prize in Economic Sciences was awarded for work directly related to volatility estimation [11].

A standard model for the covariance of asset returns is the low-rank-plus-diagonal model. It addresses many of the challenges mentioned above. At time t, this model assumes that the vector of asset returns, $x_t \in \mathbb{R}^n$, is driven by a low-dimensional vector of factor returns, $s_t \in \mathbb{R}^{n_1}$, and a vector of residual returns $\epsilon_t \in \mathbb{R}^n$, i.e.,

$$x_t = F_t s_t + \epsilon_t, \tag{2.1}$$

where $F_t \in \mathbb{R}^{n \times n_1}$ is the matrix of factor exposures, ϵ_t is independent of s_t, the n elements of ϵ_t are independent, and $n_1 \ll n$ [8]. This model implies that there are common factors that influence the returns of the assets. For equities, factors typically include the returns of specific portfolios, such as the overall market (with weights proportional to capitalization), industries, and style portfolios like the Fama-French factors [12], [13]. The model also implies that there is a part of the returns that cannot be attributed to these factors. We call this part the idiosyncratic return, ϵ_t, which is independent of the factor returns, s_t, and also independent for each asset.

Under these assumptions the asset return covariance is

$$\mathbf{cov}(x_t) = F_t \mathbf{cov}(s_t) F_t^\top + \mathbf{cov}(\epsilon_t), \tag{2.2}$$

where $\mathbf{cov}(\cdot)$ denotes the covariance. Because the elements of ϵ_t are independent, $\mathbf{cov}(\epsilon_t)$ is a diagonal matrix with positive diagonal entries. This structure is especially appealing in high dimensions (n in the order of hundreds or thousands), because we only need to estimate $\mathcal{O}(nn_1)$ parameters, instead of $\mathcal{O}(n^2)$ in the case of a generic $n \times n$ covariance matrix. The smaller number of parameters required in the low-rank-plus-diagonal structure is also a form of regularization, which avoids overfitting to the available data and can improve the covariance estimate, especially for out-of-sample applications (such as predicting future portfolio risk) [8]. The low-rank-plus-diagonal structure also allows for a positive definite covariance matrix even when there are fewer than n return samples. Finally, in the context of portfolio construction using convex optimization, the low-rank-plus-diagonal structure can be exploited to bring the computational complexity down from $\mathcal{O}(n^3)$ to $\mathcal{O}(nn_1^2)$ operations [14].

In this chapter, we assume a low-rank-plus-diagonal risk model is provided. We call this the *base* or *existing* model. This scenario is typical in the case of investors or quantitative traders, e.g., a MSCI Barra [9], [10] model is available. We propose a principled method based on maximum likelihood estimation to refine the existing model and extend it with additional statistical factors using the time series of realized asset returns. Our method can be used to capture market shifts of a shorter time horizon than the update period of the provided risk model. Furthermore, the added statistical factors capture structure in the returns that is missed by the base model. As a demonstration of our approach, we extend the Barra short-term US risk model (an already high quality and widely used in practice model) for a universe of 870 US high-capitalization equities and obtain an improvement in terms of the model's out-of-sample statistical fit. We show that the extended model uncovers hidden structure in the returns that is missed by the base model. As a secondary contribution, we also show how to handle missing return data within our framework.

The remainder of the chapter is organized as follows. In Section 2.2, we discuss related work in covariance estimation and low-rank-plus-diagonal risk modeling. In Section 2.3, we mathematically formulate our maximum likelihood estimation problem. Our proposed approach is described in Section 2.4. We discuss metrics and results in Section 2.5 and Section 2.6. The chapter concludes with Section 2.7.

2.2 *Prior work*

We split this section in two subsections. First, we review the standard approaches to empirical covariance estimation that rely solely on the time series of historical returns and do not impose any structural assumptions. Next, we turn to techniques that introduce additional structure into the covariance matrix—most notably low-rank-plus-diagonal decompositions—which aim to capture common risk factors and improve stability in high-dimensional settings.

2.2.1 *Empirical covariance estimation*

Given the time-series of historical returns, the simplest covariance estimate is the sample covariance. However, when the number of assets is large relative to the number of return samples (which is almost always true in practice), the sample covariance is a very poor approximation of the true covariance [15] and its eigenvalues can have large variance. Generally, the sample covariance may fail to capture the time-dependence and regime shifting of financial markets. A simple way to deal with the regime shifting is to use a rolling window estimate, which has a fixed memory window and only averages the return outer products in this window [8]. To improve the quality of the estimator we can add regularization or shrinkage [16]. Practitioners often replace rolling window covariance estimates with exponentially weighted moving averages (EWMAs) to place greater weight on recent observations, thereby adapting more quickly to time-varying conditions. The EWMA estimator is the smooth equivalent of the rolling window estimator and better handles the non-stationarity of financial returns [8], [10], [17]. It is controlled by the half-life parameter that dictates how far in the past the weight of an observation is reduced to 0.5. The half-life parameter is thus completely interpretable. The EWMA estimate also allows for fast, recursive computation.

Several two-stage methods exist that separately estimate the asset volatilities and correlations [8]. The estimates are then combined to obtain the final covariance estimate. In more detail, two-stage methods first form an initial covariance estimate from the return samples and then use this estimate to form transformed returns. Two-stage methods then form the final covariance estimate by using the initial covariance estimate and an estimate of the covariance of the transformed returns [8]. This procedure can be iterated to improve the quality of the covariance estimate [18]. If the initial covariance estimate is diagonal and obtained by an EWMA and the covariance of the transformed returns is also an EWMA, we get the iterated EWMA (IEWMA) predictor [8], [19]. The transformed returns are then called whitened (with volatilities close to one). The half-lives of the two EWMAs involved are typically different, with the first one (used to estimate the asset volatilities) being

smaller.

When the number of assets is large and the half-life (or the length of the returns time-series) is small, naive covariance estimation leads to singular matrices. In these cases, even though it might fail to capture information in the returns, a popular remedy is to learn a factor model [2], which we discuss next. We use the terms low-rank-plus-diagonal and factor model interchangeably, as they are equivalent in the context of risk modeling.

2.2.2 *Low-rank-plus-diagonal risk models*

Factor models are a natural way to represent correlations among assets, because assets in similar market sectors tend to behave similarly. We discuss methods that produce risk models in low-rank-plus-diagonal form. These methods differ in how they obtain the factor exposure matrix, the covariance of the factor returns, and the covariance of the idiosyncratic returns. Each method corresponds to a different category of factors: fundamental, macroeconomic, and statistical. Fundamental and macroeconomic factors are built from observable economic, financial, or asset-specific characteristics. Such a factor can correspond to a concept that has an intuitive interpretation grounded in economics, corporate structure, or investor behavior [20]. Fundamental factors include industry/sector classifications, while macroeconomic factors include style factors (e.g., value, momentum, and size). The macroeconomic Fama-French factors [12], [13] are built from returns on value-weighted or equal-weighted portfolios formed from characteristics such as size, book-to-market, profitability, or investment. Statistical factors are extracted directly from historical return data through dimensionality-reduction techniques [5], [8]. They attempt to capture the main directions of covariance in the return vector without imposing any economic interpretation.

Methods based on fundamental factors. In this case, we calculate the factor exposure matrix and then estimate the factor returns by regressing the asset returns on the factor exposures [21]. We estimate the covariance of the factor returns using a method from Section 2.2.1 and the variances of the idiosyncratic returns similarly

from the regression residuals. We mention an example of a fundamental factor: the price-to-earnings ratio. We can assign the price-to-earnings ratio of each asset as its exposure to the factor, perhaps with some cross-sectional (i.e., across-assets) standardization.

Methods based on macroeconomic factors. In this case, we calculate the factor returns first. For example, we can use the returns of style portfolios as the factor returns. We then compute the factor exposure matrix by regressing the historical asset returns on the historical factor returns [21]. We estimate the covariance of the factor returns using a method from Section 2.2.1 and the variances of the idiosyncratic returns similarly from the regression residuals. Note that the factor exposure matrix estimation is separable across assets.

Methods based on statistical factors. Such methods start by computing an empirical asset return covariance, as described in Section 2.2.1. The low-rank-plus-diagonal structure is imposed by dimensionality reduction. Effectively, we find the low-rank-plus-diagonal model that optimizes an objective involving the empirical covariance. If we minimize the Frobenius norm between the model and the empirical covariance estimate, then the solution is related to the top eigenvalues (and corresponding eigenvectors) of the empirical estimate [22], [8, Section 8.2], [23]. We can then set the idiosyncratic component of the risk model to match the diagonal entries of the empirical covariance estimate. This approach tends to underestimate the idiosyncratic risk. If we fit the risk model by maximizing the Gaussian log-likelihood, then the solution can be obtained via an iterative algorithm called expectation-maximization (EM) [8], [24], [25] or an alternating minimization approach [26]. The EM algorithm is able to cleanly handle missing data. We discuss important results related to fitting a low-rank-plus-diagonal model using the Gausssian log-likelihood objective in Chapter A. These are not important for the main flow of the chapter, but nevertheless we include them for completeness. Some ideas in Chapter A are not found in prior literature to the best of our knowledge.

We note that the risk model is time-dependent as the factor exposures, factor return covariance, and idiosyncratic return covariance are all time-dependent quantities. The update frequency of the risk model can vary from daily to, e.g., quarterly,

depending on the investor's needs and the nature of the factors involved. The update frequency can be a limiting factor when it comes to the responsiveness of the risk model to changing market conditions. For example, a quarterly update frequency may not be able to capture sudden market regime shifts or transient factors that appear on shorter time scales.

Our contribution. We propose a method that can extend any given low-rank-plus-diagonal risk model, either fundamental, macroeconomic, or statistical. Our algorithm uses the history of observed returns and a weighted Gaussian log-likelihood objective to refine the base model and extend it with additional statistical factors. As such, it is a principled way to build hybrid models with both fundamental-macroeconomic and statistical factors. Spector et al. [27] investigated how to assess the statistical fit of an existing risk model and proposed a simple method to extend the model by recursively adding factors. Our work is similarly based on a provided model, but our method both refines the given risk model and adds new statistical factors. Lee et al. [28] also propose extending a risk model by additional factors. However, the asset exposures to these new factors are not estimated statistically, but are instead constructed. The new factor exposures are thematic narrative exposures built from news data. They are constructed based on how frequently the narrative appears in news coverage about each asset. Furthermore, the covariance of the new factor returns is estimated following an empirical approach that uses ordinary least squares. In contrast, our approach relies on statistically learning the additional factors in order to improve the likelihood fit of the observed data. In addition, beyond adding new factors, our approach also refines the covariance of the factor returns of the base model.

We leave the algorithmic selection of the number of additional factors as future work. We note that our approach preserves the factor exposures to the base factors. As such, the methodology is particularly suitable when the base model consists of fundamental factors. By choosing the log-likelihood weights based on an EWMA with an appropriate (typically short) half-life, our work can also be regarded as a principled way to find additional transient factors in the returns, which are commonly known as themes [29].

2.3 Problem statement

Suppose at time t we are given a sequence of asset return vectors $x_1, \ldots, x_T \in \mathbb{R}^n$, potentially with missing entries. We are also provided with the *base* asset return covariance model

$$\mathbf{cov}(x_t) \;\approx\; F_1 \Omega_{\text{base}} F_1^\top + D_{\text{base}}, \tag{2.3}$$

where $F_1 \in \mathbb{R}^{n \times n_1}$ is the matrix of factor exposures to the *base* factors, Ω_{base} is the (positive definite) covariance of the *base* factor returns, and D_{base} is a diagonal matrix with positive diagonal entries corresponding to the idiosyncratic asset variances. In practice, the provided covariance model depends on time t, i.e., F_1, Ω_{base}, and D_{base} depend on t. For simplicity we hide this dependence.

2.3.1 The extended risk model

We consider the *extended* asset return covariance model (this model also depends on time t but we hide this dependence)

$$\mathbf{cov}(x_t) \;\approx\; \begin{bmatrix} F_1 & F_2 \end{bmatrix} \Sigma_f \begin{bmatrix} F_1^\top \\ F_2^\top \end{bmatrix} + D, \tag{2.4}$$

where $F_2 \in \mathbb{R}^{n \times n_2}$, $\Sigma_f \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$, and $D \in \mathbb{R}^{n \times n}$. D is a diagonal matrix with positive diagonal entries and Σ_f is a positive definite matrix. The number of *base* factors is n_1 and the number of added factors is n_2. Typically the total number of factors, $n_1 + n_2$, is much smaller than n. We look to find appropriate factor exposures F_2, factor return covariance Σ_f, and idiosyncratic covariance D.

We make a distinction between the factors involved in F_1 and F_2.

- F_1 is a known matrix of factor exposures, which have been measured and provided to us. For example, F_1 could be the matrix of factor exposures from an MSCI Barra model [9]. However, the covariance of the factor returns involved in F_1 will be re-estimated, by learning Σ_f. The only information from the *base* model that the *extended* model shall rely on without modification is

the matrix of factor exposures F_1. This is natural if the base model consists of fundamental factors: in such a model exposures are constructed from relatively stable asset characteristics or classification rules, while the covariance of factor returns is computed empirically.

- F_2 denotes a matrix of unknown factor exposures representing residual factor structure not captured by F_1. Estimating F_2 and Σ_f therefore allows the model to account for additional common sources of variation in asset returns beyond those included in the base model.

Identifiability of the extended risk model

Consider first a general factor model and the implied covariance $F\Sigma_f F^\top + D$, where all the terms are unknown. This is different than our setting (2.4), where F_1 is known. In the general model, neither F nor Σ_f are identifiable, since for any invertible matrix R, setting

$$\bar{F} = FR, \quad \bar{\Sigma}_f = R^{-1}\Sigma_f R^{-\top}$$

yields the same covariance. With the Cholesky decomposition $\Sigma_f = L_f L_f^\top$, another equivalent form is

$$\bar{F} = FL_f, \quad \bar{\Sigma}_f = I.$$

Since the latter is the easiest representation, the standard form for the general model assumes $\Sigma_f = I$.

Our case (2.4) is a little more complex, because F_1 is known. We show that the analogous representation for the family of risk models (2.4) is

$$\begin{bmatrix} F_1 & F_2 \end{bmatrix} \begin{bmatrix} \Omega & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} F_1^\top \\ F_2^\top \end{bmatrix} + D, \tag{2.5}$$

where $\Omega \in \mathbb{R}^{n_1 \times n_1}$ is a positive definite matrix, $F_2 \in \mathbb{R}^{n \times n_2}$ is arbitrary and D is diagonal. Note that F_2 is still only identifiable up to a rotation/reflection: if R is any $n_2 \times n_2$ orthogonal matrix, then the representation (2.5) with F_2 replaced by $F_2 R$ is equivalent, since $F_2 R R^\top F_2^\top = F_2 F_2^\top$ (a standard situation in factor models). We

prefer the representation given by the family of models (2.5), because it ensures that the additional factors remain identifiable (up to a rotation/reflection) and allows for a decomposition of risk across the *base* factors F_1 and the added factors F_2. Family (2.5) involves a smaller number of parameters compared to family (2.4), i.e., family (2.4) is over-parameterized.

Lemma 2.1. *The family of models given by the right-hand side of* (2.4) *is the same as the family of models given by* (2.5).

Proof. Consider a model $\hat{\Sigma}(F_2, \Omega, D)$ in the family (2.5). Obviously, this model also belongs to the family of models given by (2.4). Simply set

$$\Sigma_f = \begin{bmatrix} \Omega & 0 \\ 0 & I \end{bmatrix}.$$

Now, consider the model $\hat{\Sigma}\left(F_2, \Sigma_f, D\right)$ in the family (2.4) and the decomposition

$$\Sigma_f = \begin{bmatrix} \Sigma_{f,11} & \Sigma_{f,12} \\ \Sigma_{f,21} & \Sigma_{f,22} \end{bmatrix},$$

where $\Sigma_{f,11} \in \mathbb{R}^{n_1 \times n_1}$ and let

$$\Phi = \begin{bmatrix} I & \Sigma_{f,12}\Sigma_{f,22}^{-1/2} \\ 0 & \Sigma_{f,22}^{1/2} \end{bmatrix}, \quad \Omega = \Sigma_{f,11} - \Sigma_{f,12}\Sigma_{f,22}^{-1}\Sigma_{f,21}.$$

The matrix Ω is positive definite as the Schur complement of Σ_f. Because Σ_f is positive definite, $\Sigma_{f,22}$ is also positive definite.

We can write

$$\begin{aligned} \hat{\Sigma}\left(F_2, \Sigma_f, D\right) &= \begin{bmatrix} F_1 & F_2 \end{bmatrix} \Phi \begin{bmatrix} \Omega & 0 \\ 0 & I \end{bmatrix} \Phi^\top \begin{bmatrix} F_1^\top \\ F_2^\top \end{bmatrix} + D \\ &= \begin{bmatrix} F_1 & F_1\Sigma_{f,12}\Sigma_{f,22}^{-1/2} + F_2\Sigma_{f,22}^{1/2} \end{bmatrix} \begin{bmatrix} \Omega & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} F_1 & F_1\Sigma_{f,12}\Sigma_{f,22}^{-1/2} + F_2\Sigma_{f,22}^{1/2} \end{bmatrix}^\top + D. \end{aligned} \tag{2.6}$$

Therefore, the model $\hat{\Sigma}\left(F_2, \Sigma_f, D\right)$ in the family (2.4) can be represented by a model in the family (2.5). □

From (2.6), it becomes evident that the columns of the added factor exposure matrix F_2 are not necessarily orthogonal to the columns of F_1. There exists another representation that is equivalent to (2.4) and (2.5), but constrains the new factor exposure matrix F_2 to have orthogonal columns to F_1. Although not necessary for the purposes of this chapter, the lemma is given below.

Lemma 2.2. *Consider the family of risk models* (2.4). *This is equivalent to the family of risk models given by*

$$\hat{\Sigma}\left(F_{2,\perp}, \Sigma_f, D\right) = \begin{bmatrix} F_1 & F_{1,\perp} \end{bmatrix} \Sigma_f \begin{bmatrix} F_1^\top \\ F_{1,\perp}^\top \end{bmatrix} + D, \tag{2.7}$$

where $F_{1,\perp}^\top F_1 = 0$.

Proof. Consider a risk model $\hat{\Sigma}\left(F_{2,\perp}, \Sigma_f, D\right)$ in the family (2.7). It is obviously included in family (2.4).

Now consider a risk model $\hat{\Sigma}\left(F_2, \Sigma_f, D\right)$ in family (2.4). We can write

$$F_2 = F_1 U + F_{1,\perp} V,$$

where the columns of $F_{1,\perp}$ are an orthonormal basis of the orthogonal complement of the range of F_1. We can then rewrite the risk model as

$$\hat{\Sigma}\left(F_2, \Sigma_f, D\right) = \begin{bmatrix} F_1 & F_{1,\perp} \end{bmatrix} \begin{bmatrix} I & U \\ 0 & V \end{bmatrix} \Sigma_f \begin{bmatrix} I & 0 \\ U^\top & V^\top \end{bmatrix} \begin{bmatrix} F_1 & F_{1,\perp} \end{bmatrix}^\top + D$$

Therefore, the two families are the same. □

Lee et al. [28] extend a provided risk model with F_2 that has orthogonal columns to F_1. In the remainder of the chapter, we will consider the representation given by family (2.5) for the *extended* model.

2.3.2 *Estimation objective*

We assume that at each time $\tau = 1, \ldots, T$ we only observe the returns x_τ on a subset of assets indexed by the set $\mathcal{O}_\tau \subseteq \{1, \ldots, n\}$. Let $\mathcal{M}_\tau = \{1, \ldots, n\} \setminus \mathcal{O}_\tau$ be the set of assets with missing returns at time τ. We denote the observed portion of the return at τ as $x_{\tau,\text{obs}}$ and the missing portion as $x_{\tau,\text{mis}}$.

Let $p_\theta(x)$ be the probability density function of the Gaussian $\mathcal{N}(0, \tilde{F}\tilde{\Omega}\tilde{F}^\top + D)$, where

$$\tilde{F} = \begin{bmatrix} F_1 & F_2 \end{bmatrix}, \qquad \tilde{\Omega} = \begin{bmatrix} \Omega & 0 \\ 0 & I \end{bmatrix}, \tag{2.8}$$

and $\mathcal{N}(\mu, \Sigma)$ is the Gaussian distribution with mean μ and covariance Σ. The covariance of the Gaussian $\mathcal{N}(0, \tilde{F}\tilde{\Omega}\tilde{F}^\top + D)$ belongs to the family of *extended* risk models (2.5). Set the family of all components to be estimated as $\theta =: (F_2, \Omega, D)$. We estimate θ by solving

$$\underset{\theta}{\text{maximize}} \sum_{\tau \leq T} w_\tau \ell(\theta; x_{\tau,\text{obs}}), \tag{2.9}$$

where

$$\ell(\theta; x_{\tau,\text{obs}}) = \log p_\theta(x_{\tau,\text{obs}}),$$

and, with a slight abuse of notation, $p_\theta(x_{\tau,\text{obs}})$ is the marginal density of the observed returns at time τ induced by the density function $p_\theta(x)$. The weights w_τ are non-negative and sum to 1. In our setting, they reflect the assumption that more recent observations are more informative and should therefore receive larger weights. Although we do not write it explicitly, Ω is positive definite and D is diagonal with positive entries in its diagonal.

Problem (2.9) maximizes the weighted sum of the Gaussian log-likelihoods of the observed data. Although the distribution of financial returns may not be Gaussian, a Gaussian model is still a reasonable second-order approximation to the cross-sectional structure; in many risk-model applications the main object of interest is the covariance matrix, not the exact tail behavior [8]. The Gaussian assumption also gives a simple tractable framework for estimating and manipulating covariance

structure.

For most daily or even weekly equity returns, the means are typically tiny compared to the random fluctuations. Therefore, assuming a zero mean often has little practical effect when the goal is to estimate the covariance. Nevertheless, we can account for a non-zero mean, by first estimating the mean return and then subtracting it from the return samples.

2.4 Solution via expectation-maximization

We first discuss the case where all returns are observed, i.e., $\mathcal{O}_\tau = \{1, \ldots, n\}$ for all $\tau = 1, \ldots, T$. We then discuss the case of missing returns.

2.4.1 Fully observed returns

We suppose that the current time is t and we have observed the (full) return samples $x_1, \ldots, x_T \in \mathbb{R}^n$. We are also given the base risk model $F_1 \Omega_{\text{base}} F_1^\top + D_{\text{base}}$. In reality, there is time dependence in the base risk model parameters, but we drop the time index t for simplicity as before.

The following lemma holds.

Lemma 2.3. *For fully observed returns, problem* (2.9) *is equivalent* (*i.e., has the same set of solutions*) *to*

$$\underset{F_2, \Omega, D}{\text{minimize}} \quad \text{KL}\left(\mathcal{N}(0, \Sigma) \| \mathcal{N}(0, \tilde{F}\tilde{\Omega}\tilde{F}^\top + D)\right), \tag{2.10}$$

and

$$\underset{\theta}{\text{maximize}} \quad \mathbb{E}_{x \sim \mathcal{N}(0,\Sigma)} \log p_\theta(x), \tag{2.11}$$

where

$$\Sigma = \sum_{\tau \leq T} w_\tau x_\tau x_\tau^\top. \tag{2.12}$$

Proof. By the linearity of the trace operator and the fact that $\sum_{\tau \leq T} w_\tau = 1$, we get

$$\begin{aligned}
&\sum_{\tau \leq T} w_\tau \log p_\theta(x_\tau) \\
&\quad = -\frac{1}{2} \operatorname{trace}\left(\left(\tilde{F}\tilde{\Omega}\tilde{F}^\top + D \right)^{-1} \sum_{\tau \leq T} w_\tau x_\tau x_\tau^\top \right) - \frac{1}{2} \log\det\left(\tilde{F}\tilde{\Omega}\tilde{F}^\top + D \right) - \frac{1}{2} n \log(2\pi) \\
&\quad = -\frac{1}{2} \operatorname{trace}\left(\left(\tilde{F}\tilde{\Omega}\tilde{F}^\top + D \right)^{-1} \Sigma \right) - \frac{1}{2} \log\det\left(\tilde{F}\tilde{\Omega}\tilde{F}^\top + D \right) - \frac{1}{2} n \log(2\pi).
\end{aligned} \tag{2.13}$$

Problem (2.10) is equivalent (has the same solution) to

$$\operatorname*{maximize}_\theta \mathbb{E}_{x \sim \mathcal{N}(0,\Sigma)} \log p_\theta(x), \tag{2.14}$$

by expanding the KL divergence objective. By the cyclic property and the linearity of the trace operator

$$\begin{aligned}
&\mathbb{E}_{x \sim \mathcal{N}(0,\Sigma)} \log p_\theta(x) = \\
&-\frac{1}{2} \operatorname{trace}\left(\left(\tilde{F}\tilde{\Omega}\tilde{F}^\top + D \right)^{-1} \Sigma \right) - \frac{1}{2} \log\det\left(\tilde{F}\tilde{\Omega}\tilde{F}^\top + D \right) - \frac{1}{2} n \log(2\pi).
\end{aligned} \tag{2.15}$$

This proves the claim. □

Based on this, we observe that for any zero-mean distribution $\mathbb{P}$ with covariance Σ it holds that

$$\mathbb{E}_{x \sim \mathcal{N}(0,\Sigma)} \log p_\theta(x) = \mathbb{E}_{x \sim \mathbb{P}} \log p_\theta(x). \tag{2.16}$$

This justifies talking about problem (2.11) as maximum likelihood estimation, as we are maximizing the expected likelihood of the Gaussian $\mathcal{N}(0, \tilde{F}\tilde{\Omega}\tilde{F}^\top + D)$ under the generic distribution $\mathbb{P}$.

We fit (F_2, Ω, D) by solving problem (2.11) iteratively using the expectation-maximization (EM) algorithm [21]. We estimate $\theta = (F_2, \Omega, D)$ in the Gaussian factor model $x = \tilde{F}s + \varepsilon$, where $s \sim \mathcal{N}(0, \tilde{\Omega})$, $\varepsilon \sim \mathcal{N}(0, D)$, and ε is independent of

s. The density of x is $p_\theta(x)$. Given the current iterate $\theta_k = (F_2^k, \Omega_k, D_k)$, EM solves

$$\underset{\theta}{\text{maximize}}\, \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{s\sim q_k(s;x)} \log \frac{p_\theta(x,s)}{q_k(s;x)}, \tag{2.17}$$

where $q_k(s;x)$ is the conditional density of factors s given returns x under the model θ_k. The density $p_\theta(x,s)$ is the joint density of x and s under model θ. Problem (2.17) is a lower bound for problem (2.11). EM first computes

$$\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{s\sim q_k(s;x)} \log p_\theta(x,s) \tag{2.18}$$

as a function of θ, using the conditional moments of s given x under model θ_k (expectation-step). This can be done analytically using standard formulas. It then re-estimates θ by maximizing the objective (maximization-step). This yields closed form expressions.

We show how to obtain the EM update for problem (2.11) in Chapter B. The steps are outlined in Algorithm 2.1. In the case that the number of factors in F_1 is zero ($n_1 = 0$), our algorithm reduces to the standard EM algorithm [8], [24].

As demonstrated in Section B.4, the log-likelihood is (not strictly) monotone increasing with each iteration of Algorithm 2.1. Furthermore, we expect that the diagonal of the obtained risk model will be close to the diagonal of Σ from our analysis in Chapter A. When fitting a low-rank-plus-diagonal model by maximizing the Gaussian log likelihood, this is implied by the first-order optimality condition with respect to D [26, Chapter 9].

Particular attention should be given to the calculation of Σ and the initialization of the involved quantities in Algorithm 2.1. We set Σ to be an empirical covariance estimate based on the observed sample returns. We choose the exponentially weighted moving average (EWMA) covariance estimate

$$\Sigma = \alpha_T \sum_{\tau=1}^{T} \beta^{T-\tau} x_\tau x_\tau^\top, \quad \alpha_T = \left(\sum_{\tau=1}^{T} \beta^{T-\tau}\right)^{-1} = \frac{1-\beta}{1-\beta^T}, \tag{2.19}$$

where the forgetting factor β is expressed in terms of the half-life $H = -\log 2/\log\beta$.

Algorithm 2.1 EM for maximum likelihood risk model extension

Inputs: sample returns $x_1, \ldots, x_T \in \mathbb{R}^n$, base risk model $F_1 \Omega_{\text{base}} F_1^\top + D_{\text{base}}$ where $F_1 \in \mathbb{R}^{n \times n_1}$

Parameters: number of additional factors n_2, number of iterations $N_{\text{iterations}}$, half-life H, small positive threshold $\nu > 0$

Obtain Σ as an EWMA of the observed returns with forgetting factor $\beta = (1/2)^{1/H}$

▷ Initialization

Regress the returns $x_1, \ldots, x_T$ on the factor exposure matrix F_1 and obtain the residuals $\delta_1, \ldots, \delta_T \in \mathbb{R}^n$

Stack the residuals as $\Delta = \begin{bmatrix} \delta_1 & \cdots & \delta_T \end{bmatrix}$

Obtain the empirical covariance $\frac{1}{T}\Delta\Delta^\top$. Take its eigendecomposition. Let $\Lambda \in \mathbb{R}^{n_2 \times n_2}$ be the diagonal matrix of the top n_2 eigenvalues of the covariance and $U \in \mathbb{R}^{n \times n_2}$ be the matrix of the corresponding eigenvectors.

$k \leftarrow 0$

$F_2^k \leftarrow U\Lambda^{1/2}$

$\tilde{F}_k \leftarrow \begin{bmatrix} F_1 & F_2^k \end{bmatrix}$

$\Omega_k \leftarrow \Omega_{\text{base}}$

$\tilde{\Omega}_k = \begin{bmatrix} \Omega_k & 0 \\ 0 & I \end{bmatrix}$

$D_k = \max\{\text{diag}\left(\Sigma - \tilde{F}_k \tilde{\Omega}_k \tilde{F}_k^\top\right), \nu\}$, where max is elementwise and $\text{diag}(\cdot)$ with a matrix argument returns the diagonal elements

for $k = 1, \ldots, N_{\text{iterations}}$

▷ Expectation-step

$G_k^{-1} \leftarrow \tilde{F}_k^\top D_k^{-1} \tilde{F}_k + \tilde{\Omega}_k^{-1}$

$L_k \leftarrow G_k \tilde{F}_k^\top D_k^{-1}$, where $L_k = \begin{bmatrix} L_1^k \\ L_2^k \end{bmatrix}$, where $L_1^k \in \mathbb{R}^{n_1 \times n}$, $L_2^k \in \mathbb{R}^{n_2 \times n}$

$C_{ss}^k \leftarrow G_k + L_k \Sigma L_k^\top$ where $C_{ss}^k = \begin{bmatrix} C_{ss,11}^k & C_{ss,12}^k \\ C_{ss,12}^{k,\top} & C_{ss,22}^k \end{bmatrix}$ and $C_{ss,11}^k \in \mathbb{R}^{n_1 \times n_1}$, $C_{ss,12}^k \in \mathbb{R}^{n_1 \times n_2}$, $C_{ss,22}^k \in \mathbb{R}^{n_2 \times n_2}$

$\tilde{\Sigma}_k \leftarrow \Sigma + F_1 C_{ss,11}^k F_1^\top - 2\Sigma L_1^{k,\top} F_1^\top$

$C_{xs}^k \leftarrow \Sigma L_2^{k,\top} - F_1 C_{ss,12}^k$

▷ Maximization-step

$\Omega_k \leftarrow C_{ss,11}^k$

$\tilde{\Omega}_k = \begin{bmatrix} \Omega_k & 0 \\ 0 & I \end{bmatrix}$

$F_2^k \leftarrow C_{xs}^k C_{ss,22}^{k,-1}$

$\tilde{F}_k \leftarrow \begin{bmatrix} F_1 & F_2^k \end{bmatrix}$

$D_k \leftarrow \text{diag}\left(\tilde{\Sigma}_k - C_{xs}^k C_{ss,22}^{k,-1} C_{xs}^{k,\top}\right)$

return F_2^k, Ω_k, D_k

Our ability to pick up transient factors depends on the choice of the half-life parameter: a short half-life will allow us to pick up short-lived, rapidly changing effects, but if it is too short it may primarily capture noise rather than meaningful signal, whereas a longer half-life will emphasize more persistent, slowly evolving factors, but may overly smooth the data. The EWMA is not the only choice of covariance estimate. For example, we can use an IEWMA estimate. The EWMA or IEWMA estimates can be sensitive to outliers, and in this case more sophisticated approaches might be preferable.

The initialization for F_2 is based on the singular value decomposition of the residual returns of the base model. This is because the additional factors should explain the variance left unexplained by the base model. The initialization for D preserves the asset volatilities, as given by Σ. The algorithm typically converges after $20 - 30$ iterations.

2.4.2 *Missing returns*

One significant advantage of the EM approach is its ability to handle missing data. We can modify our approach to accommodate such data. We include the details of the EM approach with missing returns data in Chapter C.

2.5 *Statistical results*

We consider the time period between 2018-06-27 and 2023-12-28. This period contains the COVID-19 period that is characterized by high market volatility. Our universe consists of the US large-capitalization equities in the BlackRock Systematic investment universe for which we observe complete return data over the period 2018-06-27 and 2023-12-28. We restrict attention to assets with uninterrupted return histories and do not consider the version of our approach that explicitly accommodates missing returns for simplicity in evaluating performance: the statistical fit metrics that we consider rely on observed returns. This restriction induces a degree of survivorship bias but allows us to focus on the core mechanism by which our

approach refines a general risk model to a targeted investment universe.

Our universe consists of 870 assets. We have access to daily returns. At every time t, the returns $x_1, \ldots, x_t$ are known. The return x_τ corresponds to the time period between day $\tau - 1$ and day τ. The low-rank-plus-diagonal covariance estimate at time t is denoted

$$\hat{\Sigma}_t = \hat{F}_t \hat{F}_t^\top + \hat{D}_t. \tag{2.20}$$

Note that we hide the factor return covariance (specifically its square root) in $\hat{F}_t$. Below $\hat{\Sigma}_t$ will correspond to either

- the base model at time t, in which case $\hat{F}_t = F_{1,t}\Omega_{t,\text{base}}^{1/2}$ ($F_{1,t}$ is the provided factor exposure matrix) and $\hat{D}_t = D_{t,\text{base}}$,
- the extended model at time t, in which case $\hat{F}_t = \begin{bmatrix} F_{1,t} & \hat{F}_{2,t} \end{bmatrix} \begin{bmatrix} \hat{\Omega}_t^{1/2} & 0 \\ 0 & I \end{bmatrix}$, where $\hat{F}_{2,t}$, $\hat{\Omega}_t$, and $\hat{D}_t$ are obtained by Algorithm 2.1,
- an *exposure-misaligned* extended risk model at time t, in which case $\hat{F}_t = \begin{bmatrix} F_{1,t} & \Pi\hat{F}_{2,t} \end{bmatrix} \begin{bmatrix} \hat{\Omega}_t^{1/2} & 0 \\ 0 & I \end{bmatrix}$, where $\hat{F}_{2,t}$, $\hat{\Omega}_t$, and $\hat{D}_t$ are obtained by Algorithm 2.1 and Π is a random permutation matrix. This model serves as a baseline to show that uncovering hidden structure in the asset returns is not a trivial problem. $\Pi\hat{F}_{2,t}$ assigns the exposures to the additional factors of one asset to another asset, destroying any meaningful structure that was learned by Algorithm 2.1. We call this baseline the *extended (permuted)* model below.

Our base model is the Barra short-term US risk model, which is updated monthly. The base model consists of 73 factors for the universe of 870 assets. The 73 factors include sectors, such as aerodefense, airlines, and banks, where the exposure is binary (0 or 1), and factors where the exposures are not binary, e.g., the beta factor. We extend the base model with an additional 7 factors. This number is considered a good choice overall, as it allows the extended model to learn information missed by the base model, but also mitigates the risk of overfitting. We update the extended model on a daily basis. We use an EWMA with a half-life of $H = 126$ days as Σ

on each day. We do not optimize over the choice of the half-life or the number of additional factors in this chapter.

We compare the three models–the base, the extended, and the extended (permuted) model–with respect to their statistical fit out-of-sample. For this purpose we consider several metrics that correspond to separate subsections below. We evaluate the three models in the time period between 2019-06-26 and 2023-12-28. The evaluation period contains COVID-19. We consider the days before 2019-06-26 to be the burn-in period for the extended model.

Throughout, we assume that daily returns are zero-mean. This is the case for most daily, weekly, or monthly stock, bond, and futures returns, factor returns, and index returns [8].

2.5.1 *Log likelihood and regret*

At every day t, we compute the log likelihood of the next return, x_{t+1}, under our current model

$$l_t(\hat{\Sigma}_t) = -\frac{1}{2}\left(n\log(2\pi) + \log\det\hat{\Sigma}_t + x_{t+1}^T\hat{\Sigma}_t^{-1}x_{t+1}\right). \tag{2.21}$$

This metric is based on the assumption that the daily asset returns are Gaussian, which, as we saw earlier, is a reasonable assumption. To judge the risk models, a loss function is required and the Gaussian is a reasonable choice. A large value of log likelihood indicates a good statistical fit and suggests that the data is highly likely under our model.

Given the series of returns, the best constant predictor (in terms of log likelihood) is the empirical sample covariance

$$\Sigma^{\text{emp}} = \frac{1}{T}\sum_{t=1}^{T} x_t x_t^\top, \tag{2.22}$$

with log likelihood at time t, $l_t(\Sigma^{\text{emp}})$. Here $t = 1$ corresponds to 2019-06-27 and

$t = T$ corresponds to 2023-12-28. We therefore also report the difference

$$\Delta_t = l_t(\Sigma^{\text{emp}}) - l_t(\hat{\Sigma}_t). \tag{2.23}$$

This is a measure of regret and measures how much the covariance estimate underperforms the best possible constant covariance predictor. We want our covariance estimate to have small regret. Note that the regret can be negative, because our covariance is time-dependent, i.e., a time-dependent covariance can outperform Σ^{emp}.

Our results are included in Table 2.1. We show both the average log likelihood and regret over the evaluation period, but also the standard deviation of these averages. The extended model obtains a lift in both log-likelihood and regret. The extended (permuted) model is worse than the base model.

Model	Average Log Likelihood	Std. of Average Log-Likelihood	Average Regret	Std. of Average Regret
Base	2330.53	33.77	491.38	33.15
Extended	**2371.73**	27.99	**450.18**	27.30
Extended (permuted)	2273.73	38.59	548.18	38.01

Table 2.1: Log likelihood and log likelihood regret metrics.

2.5.2 *Whitened returns*

For each risk model, we compute out-of-sample whitened returns and their empirical correlation matrix. We judge the risk model by the discrepancy between the empirical correlation matrix and the implied (by the risk model) correlation matrix of the whitened returns.

In detail, consider an arbitrary risk model $\hat{\Sigma}_t$ for every time t. Under the hypothesis that the risk model is true and the returns are zero-mean, then $\mathbf{cov}(x_t) = \hat{\Sigma}_t$ and $\mathbf{cov}(\hat{\Sigma}_t^{-1/2} x_t) = I$, where I is the identity.

We now collect the out-of-sample whitened returns $\{\hat{\Sigma}_t^{-1/2} x_{t+1}\}_{t=1}^{T}$ that should have the identity as the covariance under this hypothesis. We compute the Frobenius

norm between the empirical correlation matrix ($n \times n$ matrix) of the whitened returns $\{\hat{\Sigma}_t^{-1/2} x_{t+1}\}_{t=1}^T$ and their theoretical correlation matrix assuming the risk model is true, i.e., the identity matrix I. We report the Frobenius norm for each risk model in Table 2.2. A larger normalized Frobenius norm indicates a greater discrepancy between the risk model and the covariance structure of the realized returns. The extended model achieves the best fit.

Model	Frobenius norm of correlation matrix to identity
Base	67.15
Extended	**49.20**
Extended (permuted)	68.36

Table 2.2: Results for the whitened returns.

2.5.3 *Whitened residuals*

We perform a similar procedure to the previous section, but instead of looking at out-of-sample whitened returns, we focus on out-of-sample whitened residuals. We describe our approach using an arbitrary risk model $\hat{\Sigma}_t$ for every time t.

At time t, we solve the out-of-sample ordinary least squares problem

$$\underset{s_t}{\text{minimize}} \|x_{t+1} - \hat{F}_t s_t\|_2^2 \tag{2.24}$$

and compute the residual

$$\hat{\delta}_t = x_{t+1} - \hat{F}_t \hat{s}_t \tag{2.25}$$

where $\hat{s}_t$ is the solution of (2.24).

We assume that the risk model $\hat{\Sigma}_t$ holds for the returns, i.e., the returns follow

$$x_t = \hat{F}_t s_t + \epsilon_t, \quad \mathbf{cov}(s_t) = I, \quad \mathbf{cov}(\epsilon_t) = \hat{D}_t.$$

By letting $M_t = I - \hat{F}_t(\hat{F}_t^\top \hat{F}_t)^{-1}\hat{F}_t^\top$, it also holds that the in-sample ordinary least squares residuals e_t (where we solve (2.24) but replacing x_{t+1} with x_t) are equal to

$M_t\epsilon_t$. Since M_t is an orthogonal projection matrix it can be written as $M_t = U_t U_t^\top$, where U_t is orthonormal. Let $z_t = U_t^\top e_t$. Then $\mathbf{cov}(z_t) = U_t^\top \hat{D}_t U_t$ and

$$\mathbf{cov}\left((U_t^\top \hat{D}_t U_t)^{-1/2} z_t\right) = I. \tag{2.26}$$

We collect the out-of-sample whitened residuals

$$\{(U_t^\top \hat{D}_t U_t)^{-1/2} U_t^\top \hat{\delta}_t\}_{t=1}^T \tag{2.27}$$

and compute the Frobenius norm between the empirical correlation matrix of the out-of-sample whitened residuals and their theoretical correlation matrix under the assumption that the risk model holds, i.e., the identity I. For each of the risk models, we report the Frobenius norm in Table 2.3. A larger normalized Frobenius norm indicates a greater discrepancy between the risk model and the covariance structure of the realized returns. The extended model achieves the best statistical fit.

Model	Frobenius norm of correlation matrix to identity
Base	62.80
Extended	**43.65**
Extended (permuted)	49.86

Table 2.3: Results for the residuals.

2.5.4 Out-of-sample R^2

If the extended model uncovers covariance structure that the base model misses, then it should be able to better estimate out-of-sample returns. In other words, the extended model should explain more of the out-of-sample returns variance. We measure this by computing the out-of-sample R^2. We analyze the R^2 obtained from estimating returns and the R^2 obtained from estimating the residuals relative to the base factors, recognizing that these quantities are closely related. A higher R^2 implies a better statistical fit. A negative R^2 implies that the model is worse than

simply predicting zero.

Out-of-sample R^2 for the returns

At time t, we split the n assets in two disjoint groups G_{train} and G_{test} (we suppress the dependence on t for these two groups for clarity). We solve the out-of-sample maximum likelihood problem tied with the risk model (2.20), but limited to the set of assets G_{train}:

$$\underset{s_t}{\text{minimize}} \left(x_{t+1}^{\text{train}} - \hat{F}_t^{\text{train}} s_t\right)^\top \hat{D}_t^{\text{train},-1} \left(x_{t+1}^{\text{train}} - \hat{F}_t^{\text{train}} s_t\right) + s_t^\top s_t, \tag{2.28}$$

where the superscript train indicates that we only consider the assets in G_{train}. In other words, our estimate of the factor returns is the conditional expectation under our factor model given the returns in train. We then compute the predicted returns for the assets in G_{test} as

$$\hat{x}_{t+1}^{\text{test}} = \hat{F}_t^{\text{test}} s_t^\star, \tag{2.29}$$

where $s_t{}^\star$ is the solution of (2.28). This is the conditional expectation under our factor model given the returns in train. Finally, we compute the R^2 between the predicted next-day returns and the actual next-day returns for the assets in G_{test}:

$$R_t^2 = 1 - \frac{\|x_{t+1}^{\text{test}} - \hat{x}_{t+1}^{\text{test}}\|_2^2}{\|x_{t+1}^{\text{test}} - \mathbf{0}\|_2^2}. \tag{2.30}$$

Using zero as the mean is reasonable for asset returns. A higher R^2 implies that the risk model captures more of the structure present in the returns, as it is able to better estimate out-of-sample returns. We compute R_t^2 for every day in the evaluation period.

For each of 30 replications, at each time t we randomly split the assets into a test and train set (90% of the assets are in the train set and the remaining 10% is in the test set at each time) and compute the corresponding out-of-sample return R_t^2. We then average R_t^2 across replications at each t to obtain a cross-replication mean time series. This time series is shown in Figure 2.2. The extended model

consistently captures covariance structure that the base model misses, resulting in a higher out-of-sample return R^2. Nevertheless, the base model also captures a significant amount of structure in the returns. The extended (permuted) model performs significantly worse: assigning the added factors' exposures of one asset to another is enough to wash out the covariance structure uncovered by Algorithm 2.1. This indicates that uncovering hidden structure in the asset returns is not a trivial problem. The extended (permuted) model is significantly worse at the end of 2023. This is consistent with a significant regime shift that occurred in the US equity market during the final quarter of 2023. The transition was from a narrow market led only by a few technology giants to a more broad-based rally. In this period, misaligning exposures across assets leads to a significant degradation in performance.

The 90/10 split is chosen to ensure that the train set is large enough to obtain meaningful factor return estimates, while the test set is large enough to obtain meaningful out-of-sample R^2 estimates. We have verified that other splits (e.g., 50/50) lead to similar conclusions.

For each risk model we report a point estimate and a dispersion for the R^2. The reported point estimate corresponds to the time average of the cross-replication mean. The reported dispersion measure is the standard deviation of the average across time of the cross-replication mean. These results are included in Table 2.5. The extended model offers an improvement in R^2.

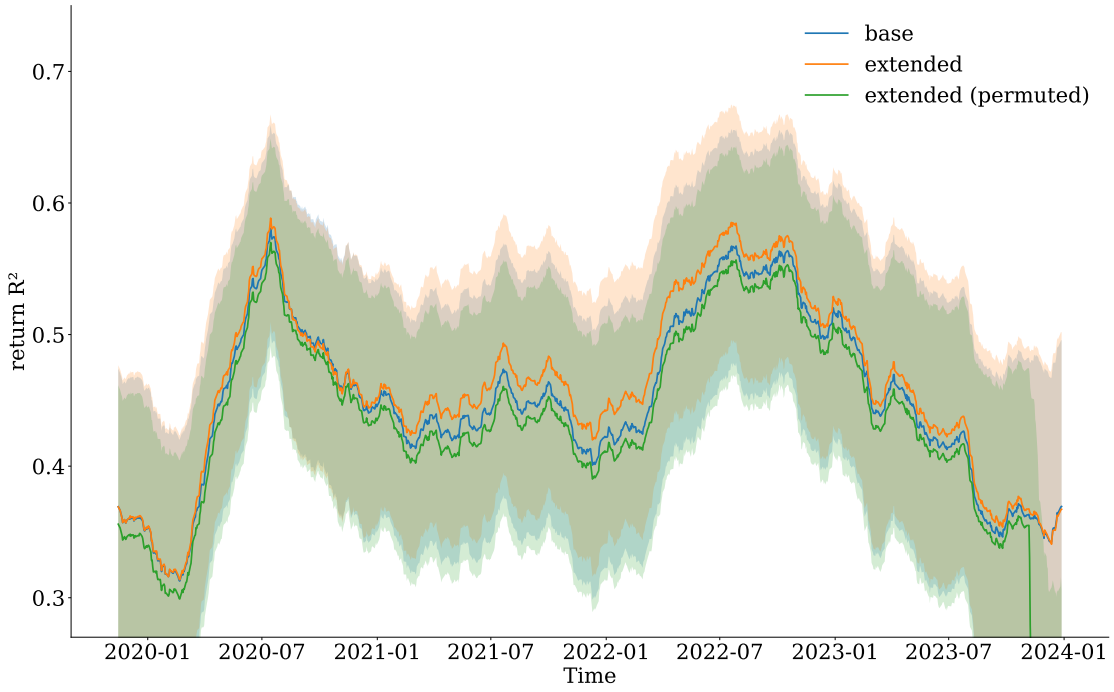


Figure 2.1: Out-of-sample return predictability (R^2) across time. At each of 30 replications, assets are randomly split into train and test sets (90/10) at each time, and the out-of-sample (next-day for test assets) return R^2 is computed for each day. Lines report the rolling mean (window = 100 days) of the cross-replication average R^2. The shaded bands indicate the ± 1 rolling mean of the cross-replication standard deviation of R^2 (window = 100 days), providing a smoothed measure of sampling variability across random splits.

Model	Point estimate	Dispersion
Base	0.445	0.006
Extended	**0.454**	0.006
Extended (permuted)	0.363	0.02

Table 2.4: Results for the out-of-sample return R^2.

Out-of-sample R^2 for the base residuals

We repeat the same procedure as above, but now we compute the R^2 for estimating the residuals with respect to the base factors, which we call the *base* residuals. We check whether the extended model is able to capture structure in the out-of-sample base residuals that is missed by the base model. This methodology is similar, but not identical, to the measure of out-of-sample performance proposed by Spector et al. [27].

At each time t, we split the n assets in two disjoint groups G_{train} and G_{test} (again we suppress the dependence on t).

- **Extended model**. In this case, $\hat{F}_t = \begin{bmatrix} \hat{F}_{1,t} & \hat{F}_{2,t} \end{bmatrix}$, where $\hat{F}_{1,t} = F_{1,t}\hat{\Omega}_t^{1/2}$ corresponds to the n_1 base factors and $\hat{F}_{2,t}$ corresponds to the added n_2 factors. We first compute the G_{train} residuals with respect to the base factors by solving

$$\begin{aligned} &\underset{s_t^{(1)}}{\text{minimize}}\ s_t^{(1),\top} s_t^{(1)} + \\ &\left(x_{t+1}^{\text{train}} - \hat{F}_{1,t} s_t^{(1)}\right)^\top \left(\hat{F}_{2,t}\hat{F}_{2,t}^\top + \hat{D}_t^{\text{train}}\right)^{-1} \left(x_{t+1}^{\text{train}} - \hat{F}_{1,t} s_t^{(1)}\right), \end{aligned} \tag{2.31}$$

  where the superscript train indicates that we only consider the assets in G_{train}. The G_{train} residuals with respect to the base factors are then:

$$\epsilon_{t+1}^{\text{train}} = x_{t+1}^{\text{train}} - \hat{F}_{1,t}^{\text{train}} s_t^{(1),\star}, \tag{2.32}$$

  where $s_t^{(1),\star}$ is the solution of (2.31). In other words, we use the conditional expectation for the base factor returns under our factor model. Problem (2.31) gives just that because the remaining term in the factor model, excluding the term with the base factors has covariance $\hat{F}_{2,t}\hat{F}_{2,t}^\top + \hat{D}_t$ and is independent of the base factors.

  We then estimate the added factors' returns by regressing the base residuals

on the added factors' exposures:

$$\underset{s_t^{(2)}}{\text{minimize}} \left(\epsilon_{t+1}^{\text{train}} - \hat{F}_{2,t}^{\text{train}} s_t^{(2)}\right)^\top \hat{D}_t^{\text{train},-1} \left(\epsilon_{t+1}^{\text{train}} - \hat{F}_{2,t}^{\text{train}} s_t^{(2)}\right) + s_t^{(2),\top} s_t^{(2)}, \tag{2.33}$$

with solution $s_t^{(2),\star}$.

We now compute the residuals for the assets in G_{test} with respect to the base factors by solving

$$\underset{s_t^{(1)}}{\text{minimize}} \left(x_{t+1}^{\text{test}} - \hat{F}_{1,t}^{\text{test}} s_t^{(1)}\right)^\top \left(\hat{F}_{2,t}^{\text{test}} \hat{F}_{2,t}^{\text{test},\top} + \hat{D}_t^{\text{test}}\right)^{-1} \left(x_{t+1}^{\text{test}} - \hat{F}_{1,t}^{\text{test}} s_t^{(1)}\right) + s_t^{(1),\top} s_t^{(1)}, \tag{2.34}$$

and setting

$$\epsilon_{t+1}^{\text{test}} = x_{t+1}^{\text{test}} - \hat{F}_{1,t}^{\text{test}} \tilde{s}_t^{(1),\star}, \tag{2.35}$$

where $\tilde{s}_t^{(1),\star}$ is the solution of (2.34).

Finally, we compute the predicted base residuals for the assets in G_{test} using the added factors as

$$\hat{\epsilon}_{t+1}^{\text{test}} = \hat{F}_{2,t}^{\text{test}} s_t^{(2),\star}, \tag{2.36}$$

and find the R^2 between the predicted next-day base residuals and the actual next-day base residuals for the assets in G_{test}

$$R_t^2 = 1 - \frac{\|\epsilon_{t+1}^{\text{test}} - \hat{\epsilon}_{t+1}^{\text{test}}\|_2^2}{\|\epsilon_{t+1}^{\text{test}} - \mathbf{0}\|_2^2}. \tag{2.37}$$

A positive R^2 here indicates that the added factors are able to explain existing structure in the out-of-sample residuals that is missed by the base factors. We note that the predicted base residuals for G_{test} only depend on the next day's returns of the assets in G_{train}, while the base residuals for G_{test} only depend on the next day's returns of the assets in G_{test}. Our ability to predict the latter using the former depends on the quality of the risk model.

- **Base model**. In this case $\hat{F}_t = \left[F_{1,t}\Omega_{t,\text{base}}^{1/2}\right]$, where $F_{1,t}$ corresponds to the provided base factor exposures and $\Omega_{t,\text{base}}$ corresponds to the provided base factor covariance at time t, i.e., $F_{1,t}$ corresponds to F_1 and Ω_t corresponds to Ω_{base} of Section 2.4.

  We compute the residuals on G_{test} by solving (2.34) and using (2.35) where we replace $\hat{F}_{1,t}$ with $F_{1,t}\Omega_{t,\text{base}}^{1/2}$ and $\hat{F}_{2,t}$ with $\mathbf{0}$. We then compute the R^2 between the predicted base residuals and the actual residuals for the assets in G_{test} as

$$R_t^2 = 1 - \frac{\|e_{t+1}^{\text{test}} - \mathbf{0}\|_2^2}{\|e_{t+1}^{\text{test}} - \mathbf{0}\|_2^2} = 0. \tag{2.38}$$

  Our prediction is $\mathbf{0}$ because there are no added factors to explain the base residuals in the base model.

A positive R^2 for the extended model (2.37) implies that it captures additional structure present in the base residuals. The R_t^2 (2.38) for the base model should be 0 as there is no assumed structure in the base residuals for this model. We compute R_t^2 for every day in the evaluation period.

For each of 30 replications, at each time t we randomly split the assets into a test and a train set (the train set contains 90% of the assets and the test set the remaining 10%) and compute the corresponding out-of-sample residual R_t^2 at each time t. For the extended model and the extended (permuted) model we use eq. (2.37), while for the base model we use eq. (2.38). We then average R_t^2 across replications at each t to obtain a cross-replication mean time series. This time series is shown in Figure 2.2. The extended model consistently captures residual structure that the base model misses, resulting in a positive out-of-sample residual R^2. The extended (permuted) model obtains an R^2 that is negative. This implies that a wrong exposure attribution for the added factors leads to a model that is worse than simply predicting zero for the out-of-sample base residuals. The behavior of the extended (permuted) model at the end of 2023 is in line with our observations from Figure 2.2.

For each risk model we report a point estimate and a dispersion for the R^2. The reported point estimate corresponds to the time average of the cross-replication

mean. The reported dispersion measure is the standard deviation of the average across time of the cross-replication mean. These results are included in Table 2.5. The extended model offers an improvement in R^2.

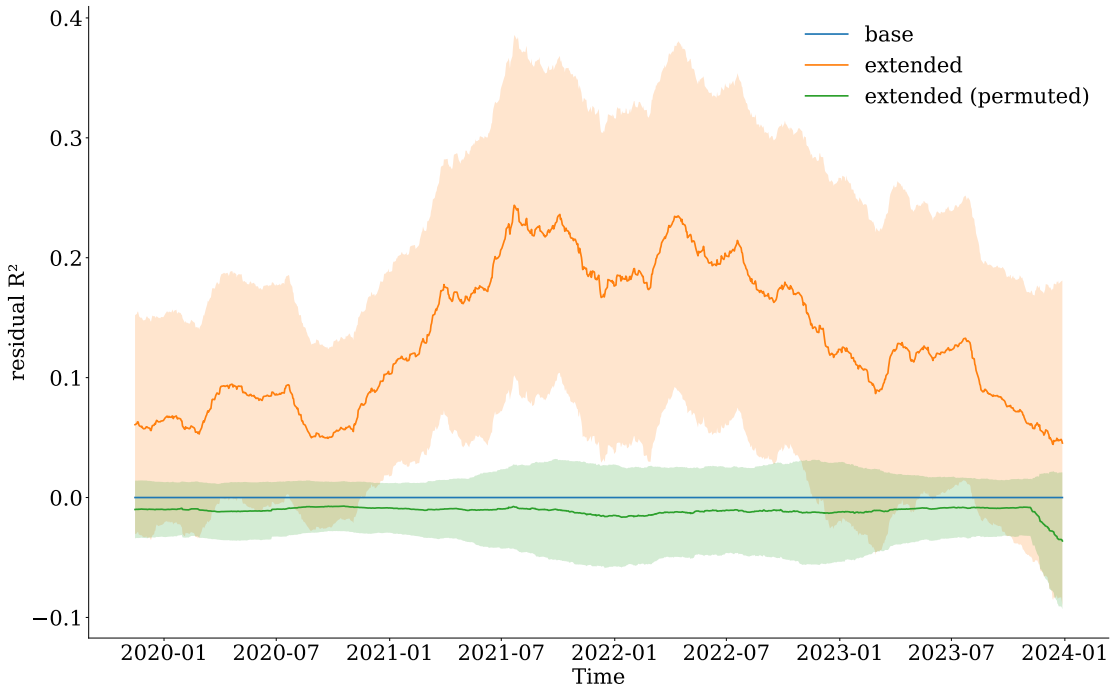


Figure 2.2: Out-of-sample predictability (R^2) of the residuals with respect to the base model's factors across time. At each of 30 replications and at each time, assets are randomly split into train and test sets (90/10), and the out-of-sample (next-day for test assets) R^2 of the base residuals is computed for each day. Lines report the rolling mean (window = 100 days) of the cross-replication average R^2. The shaded bands indicate the ± 1 rolling mean of the cross-replication standard deviation of R^2 (window = 100 days), providing a smoothed measure of sampling variability across random splits.

2.5.5 *Predictive ability of added factors*

We demonstrate that the added factors of the extended model capture structure present in the returns that the base factors miss. In this section we consider the

Model	Point estimate	Dispersion
Base	0.0	0.0
Extended	**0.125**	0.005
Extended (permuted)	−0.013	0.001

Table 2.5: Results for the out-of-sample residual R^2.

extended risk model $\hat{F}_t = \begin{bmatrix} \hat{F}_{1,t} & \hat{F}_{2,t} \end{bmatrix}$, where $\hat{F}_{1,t} = F_{1,t}\hat{\Omega}_t^{1/2}$ corresponds to the n_1 base factors and $\hat{F}_{2,t}$ corresponds to the added n_2 factors.

We consider the null hypothesis

$$x_t = \hat{F}_{1,t} s_t^{(1)} + \tilde{\epsilon}_t, \quad (H_0) \tag{2.39}$$

where $s_t^{(1)}, \tilde{\epsilon}_t$ are independent, and the elements of $\tilde{\epsilon}_t$ are independent. In other words, we assume there is no added structure, other than the base factors. We will show that our predictive ability is not consistent with this hypothesis.

We split the n assets in two disjoint groups G_{train} and G_{test} (again we suppress the dependence on t) and solve the out-of-sample ordinary least squares problems

$$\underset{s_t}{\text{minimize}} \|x_{t+1}^{\text{train}} - \hat{F}_{1,t}^{\text{train}} s_t\|_2^2, \tag{2.40}$$

$$\underset{s_t}{\text{minimize}} \|x_{t+1}^{\text{test}} - \hat{F}_{1,t}^{\text{test}} s_t\|_2^2. \tag{2.41}$$

We compute the optimal residuals

$$\hat{\delta}_t^{\text{train}} = x_{t+1}^{\text{train}} - \hat{F}_{1,t}^{\text{train}} {s_t}^{\star,\text{train}} \tag{2.42}$$

and

$$\hat{\delta}_t^{\text{test}} = x_{t+1}^{\text{test}} - \hat{F}_{1,t}^{\text{test}} {s_t}^{\star,\text{test}} \tag{2.43}$$

where ${s_t}^{\star,\text{train}}$ and ${s_t}^{\star,\text{test}}$ are the solutions of (2.40) and (2.41), respectively. Under the null, it holds that $\hat{\delta}_t^{\text{train}}$ and $\hat{\delta}_t^{\text{test}}$ are independent. In other words, under the null the residuals $\hat{\delta}_t^{\text{train}}$ are not predictive of the residuals $\hat{\delta}_t^{\text{test}}$: the highest possible R^2 value is 0.

Now let

$$P_t^{\text{train}} = \left(I - \hat{F}_{1,t}^{\text{train}}(\hat{F}_{1,t}^{\text{train},\top}\hat{F}_{1,t}^{\text{train}})^{-1}\hat{F}_{1,t}^{\text{train},\top}\right)\hat{F}_{2,t}^{\text{train}} \tag{2.44}$$

$$P_t^{\text{test}} = \left(I - \hat{F}_{1,t}^{\text{test}}(\hat{F}_{1,t}^{\text{test},\top}\hat{F}_{1,t}^{\text{test}})^{-1}\hat{F}_{1,t}^{\text{test},\top}\right)\hat{F}_{2,t}^{\text{test}}. \tag{2.45}$$

Under the alternative hypothesis that the extended risk model holds, i.e., the additional factors exist and

$$x_t = \hat{F}_{1,t}s_t^{(1)} + \hat{F}_{2,t}s_t^{(2)} + \epsilon_t, \quad (H_1), \tag{2.46}$$

the in-sample residuals of the train and test assets with respect to the base factors have factor exposures to $s_t^{(2)}$ equal to P_t^{train} and P_t^{test}, respectively. We estimate $s_t^{(2),\star}$ by solving

$$\underset{s_t}{\text{minimize}}\|\hat{\delta}_t^{\text{train}} - P_t^{\text{train}}s_t\|_2^2. \tag{2.47}$$

and make the prediction

$$\tilde{\delta}_t^{\text{test}} = P_t^{\text{test}}s_t^{(2),\star} \tag{2.48}$$

for $\hat{\delta}_t^{\text{test}}$. Under the null, the best prediction would be $\mathbf{0}$ for $\hat{\delta}_t^{\text{test}}$ (assume $\hat{F}_{2,t} = 0$) and the max R^2 would also be 0. If the R^2

$$R_t^2 = 1 - \frac{\|\hat{\delta}_t^{\text{test}} - \tilde{\delta}_t^{\text{test}}\|_2^2}{\|\hat{\delta}_t^{\text{test}} - \mathbf{0}\|_2^2} \tag{2.49}$$

is positive, then we are able to predict $\hat{\delta}_t^{\text{test}}$ from $\hat{\delta}_t^{\text{train}}$, which should not be possible under the null. Therefore, this would be evidence against the null and in favor of the discovered added factors. This methodology is inspired by Spector et al. [27]. We call this the mosaic test inspired by the title of their chapter [27].

For each of 30 replications, at each time t we randomly split the assets into a test and a train set (the train set contains 90% of the assets and the test set the remaining 10%) and compute the corresponding out-of-sample residual R_t^2 (2.49) at each time t for the extended and extended (permuted) risk models. We then average R_t^2 across replications at each t to obtain a cross-replication mean time series. This time series is shown in Figure 2.3. The predictive ability implied by the positive

out-of-sample R^2 for the extended model is evidence that the added factors are present (equivalently evidence against the null).

We also carry out the procedure for the extended (permuted) model. The extended (permuted) model obtains an R^2 that is negative. This implies that a wrong exposure attribution for the added factors does not allow us to reject the null hypothesis of no added factors with this model.

For each risk model we report a point estimate and a dispersion for the R^2. The reported point estimate corresponds to the time average of the cross-replication mean. The reported dispersion measure is the standard deviation of the average across time of the cross-replication mean. These results are included in Table 2.5. The extended model offers an improvement in R^2.

Model	Point estimate	Dispersion
Extended	**0.0125**	0.002
Extended (permuted)	-0.0168	0.0005

Table 2.6: Evidence (R^2) against the null 'no added factors'. A positive value indicates the existence of added factors in the returns.

2.6 *Portfolio performance*

We also evaluate the performance of the risk models in a portfolio optimization task. We consider a long-short portfolio optimization problem, similar to the ones solved by most hedge-funds [5]. At every time t that trades are calculated, our optimization problem involves an alpha term, $\mathbb{E}[r_t]^\top w$, transaction costs, and borrow costs, and a constraint involving the risk model available at time t that limits the portfolio volatility. We also include a turn-over limit, weight limits, and trade limits.

We perform the same backtest over the same time period for each family of risk models. For the portfolio corresponding to the extended model, we use the covariance matrix implied by the extended risk model at each time t in our optimization problem. For the portfolio corresponding to the base model, we use the covariance matrix implied by the base risk model at each time t.

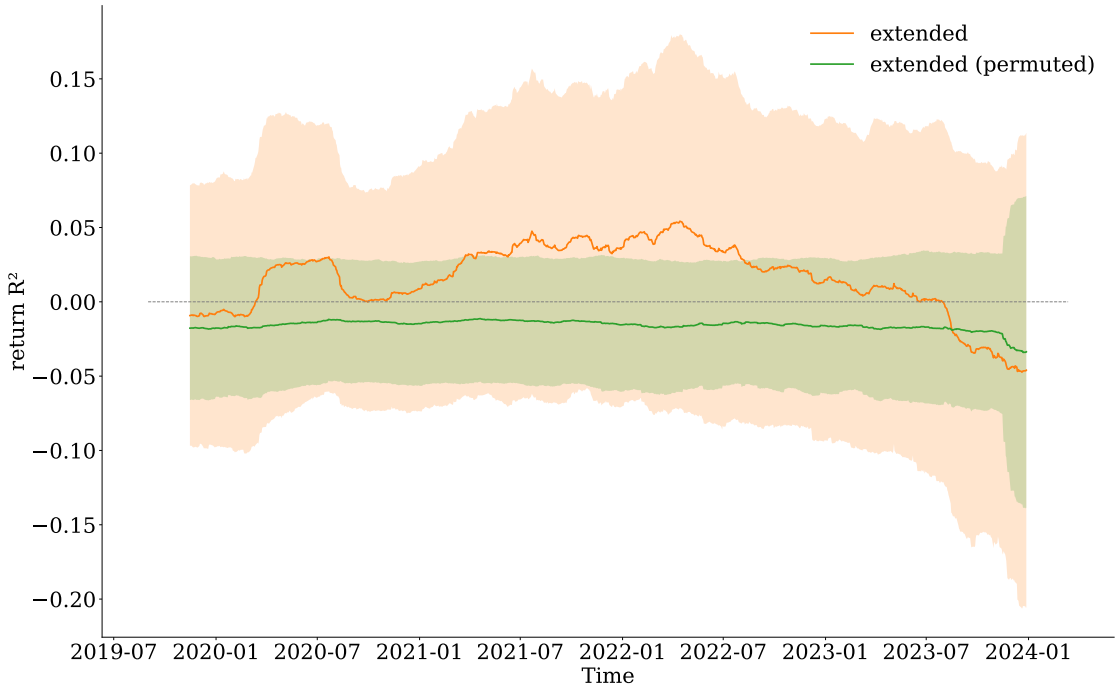


Figure 2.3: Evidence against the null 'no added factors' (R^2) for the extended and extended (permuted) risk models. At each of 30 replications and at each time, assets are randomly split into train and test sets (90/10), and the out-of-sample R^2 (2.49) is computed for each day. Lines report the rolling mean (window = 100 days) of the cross-replication average R^2. The shaded bands indicate the ± 1 rolling mean of the cross-replication standard deviation of R^2 (window = 100 days), providing a smoothed measure of sampling variability across random splits.

The Pareto frontier for the two portfolios is shown in Figure 2.4. The extended model's portfolio dominates the base model's portfolio, as it achieves higher returns for the same level of realized risk. This is consistent with our previous results showing that the extended model captures more structure in the returns than the base model.

Nevertheless, the performance of the risk models in the portfolio optimization task should not be taken as a direct consequence of the statistical fit results. A statistically better risk model need not perform better for portfolio optimization, as the latter depends on the specific optimization problem and the quality of the alpha

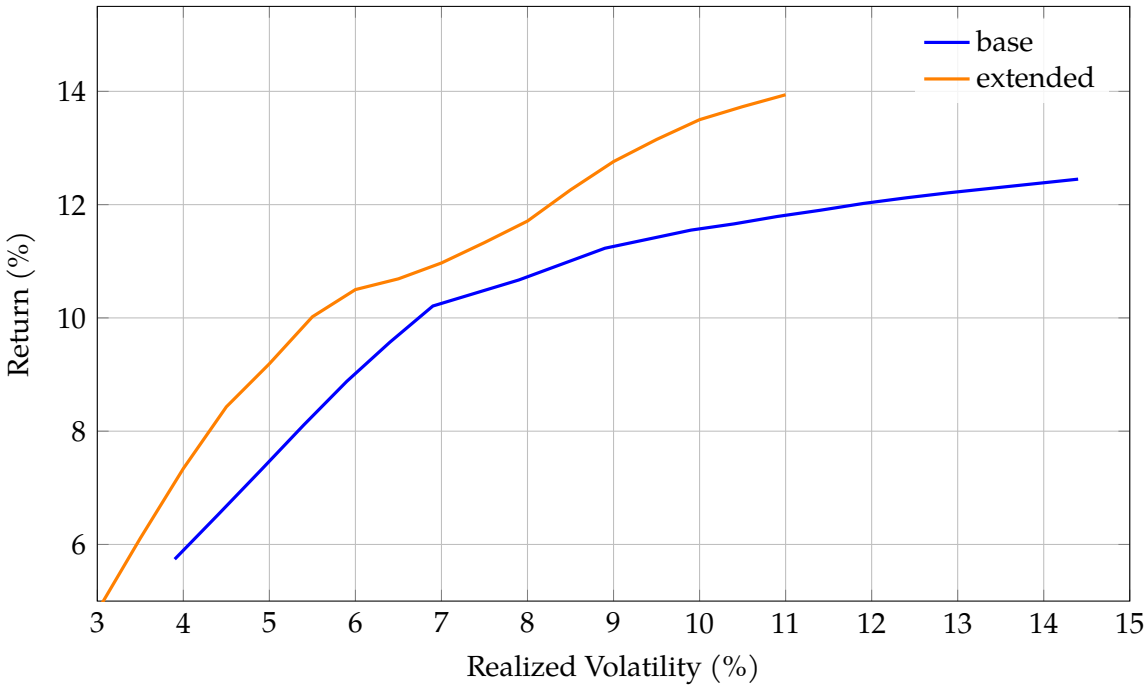


Figure 2.4: Risk–return tradeoff as a function of realized volatility for the base and extended risk models.

signal. There are cases where a bad alpha signal can lead to a better performance when a statistically worse risk model is used in the optimization. This happens for example if the risk model over-estimates risk and overly constrains any trades, which can be beneficial if the alpha signal is very bad.

2.7 *Discussion*

We propose a methodology to refine a provided low-rank-plus-diagonal risk model and extend it with additional statistical factors. Our method is based on maximum likelihood estimation and is an instance of the expectation-maximization algorithm. Our method is particularly useful if the provided risk model is only infrequently updated, but we would like to incorporate information in shorter time horizons as well. We empirically show that this modification can improve the risk model's statistical fit even when the provided risk model is of high quality. Our approach is

also able to handle missing returns data.

However, we note that adding factors does not necessarily improve statistical fit [27]. Furthermore, the effectiveness of our approach depends on the quality of the factors in the original risk model, namely the quality of the provided factor exposure matrix.

We leave the problem of selecting the number of additional factors as future work, although we do note that a value in the range 2-10 typically worked well. A simple approach to selecting the number of factors would be by looking at the out-of-sample R^2.

In the future, we plan to combine our extension methodology with factor selection in the base model. This can be helpful if the base model has many factors, but only a few of them are relevant. We can reduce the number of base factors using, e.g., forward selection, and then apply our extension methodology to the reduced base model. We can also consider learning an F_2 with a specified range, if we wish to capture specific factors.

Deriving the EM update for a block-diagonal matrix D would be useful for extending a factor model that captures both equities and exchange traded funds (ETFs). In addition, adding a penalty that encourages sparsity in the precision matrix in our objective can be useful as a form of regularization.

Future work will also deal with interpreting the added statistical factors. This can be done by looking at the cross-sectional correlation of the added factors with existing themes or by querying a large language model.

3 *Geometric representation*

In the previous chapter, we discussed a method to estimate a covariance model. We studied the problem of portfolio optimization where the covariance is sufficient to capture portfolio risk. Many control problems however require more granular information of uncertainty compared to what the covariance provides, e.g., cases with multi-modal uncertainty or important tail events. The extreme case is a control problem that depends on the distribution of uncertainty. For example, suppose we wish to steer the distribution of the state of a dynamical system to a target distribution. Scalable ways to compute the similarity between two probability distributions then become important.

In this chapter, we introduce a representation of uncertainty based on one-dimensional projections that allows to efficiently compare two high-dimensional distributions. As such, the representation can be helpful for control problems that involve the distribution of uncertainty. We call this representation *geometric* because it is based on Euclidean projections.

The *geometric* representation is a family of distances between the probability distributions of two random variables that is based on the discrepancy between the cumulative distribution functions of random linear one-dimensional projections of the random variables. Our proposed distance is interpretable, computationally simple, and admits a differentiable approximation. We establish asymptotic theoretical guarantees for sample-based estimators of the distance. We empirically study the use of the distance in two-sample tests and demonstrate its ability to distinguish different distributions. Finally, we show that the distance allows for simple gradient-based solutions in control by studying distribution steering and

ergodic control.

3.1 Introduction

Quantifying how far two probability distributions are from one another is a central task in many sequential decision making applications, such as distribution steering [30], [31], [32], and ergodic control [33], [34]. In distribution steering, the Kullback–Leibler divergence [30], the Wasserstein distance [31], or the difference between the moments of the distributions [32] is often used. In ergodic control, the discrepancy between the Fourier coefficients of the two distributions has been suggested [34], [35], while the distance from the simplex center has been explored in sequential problems [36].

For the case of distributions over the reals, various notions of similarity have been proposed, such as divergences [37], the Wasserstein metric [38], the Kolmogorov distance [39], and the continuous ranked probability score (CRPS) [40]. Many of these (i.e., the first three) extend to higher dimensions, but become computationally heavy. The last one cannot be easily extended to dimensions larger than one. To compute distances between high-dimensional distributions, in the machine learning literature [41], [42], projection-based methods have been proposed. These methods average discrepancies between one-dimensional projections of the distributions and therefore are typically called *sliced*. Any of the aforementioned discrepancies for distributions over the reals can then be used for the discrepancy between the one-dimensional projections. To the best of our knowledge, sliced techniques are not commonly used in the area of control. This is because random projections make safety guarantees harder and the variance of sliced distances can be large, which can lead to poor performance.

In this chapter, we unify prior work in sliced distances and introduce a family of sliced distances between two distributions that is based on the cumulative distribution functions (CDFs) of their linear one-dimensional projections. Specifically, we measure the expected dissimilarity between the CDFs of their one-dimensional linear projections. We allow for flexibility in choosing the dissimilarity function; various sliced distances are members of our

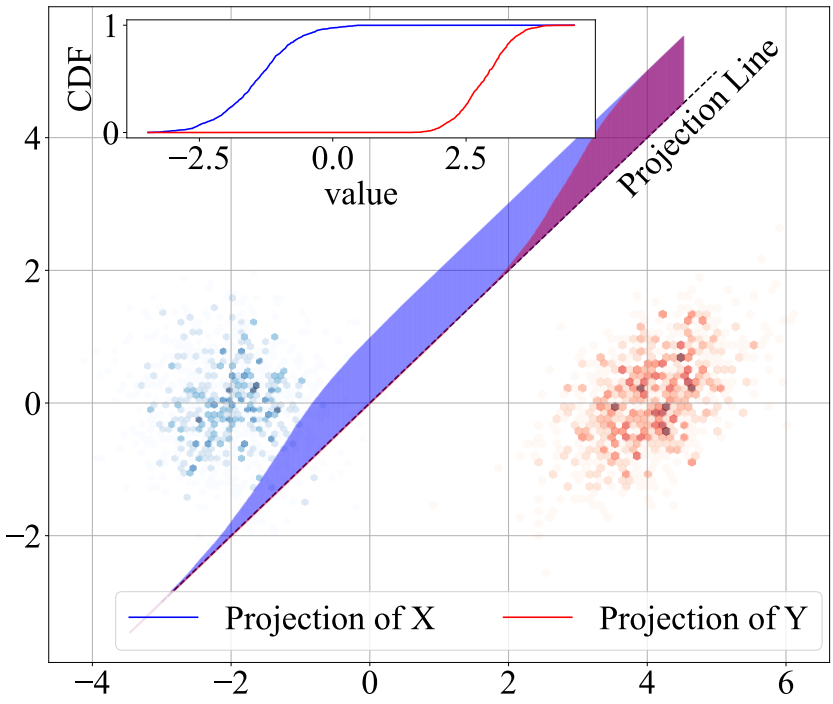


Figure 3.1: Overview of our proposed family of distances between the distributions of two random variables X and Y. We average discrepancies between the CDFs of linear one-dimensional projections of the random variables. The distributions of the random variables are represented as sets of samples here. We only depict a single projection (and the corresponding CDF of the projection of each random variable) for illustration purposes, but the distance is computed by averaging over many projections.

proposed family, as discussed below. The distances in our introduced family are interpretable and a differentiable estimate can be found using samples from the two distributions; we simply average (functions of) the difference in probability content over chosen half-spaces for the two distributions. The main elements of our proposed distance are summarized in Figure 3.1.

Su et al. [43] proposed L^p distances between CDFs for the space of one-dimensional distributions. However, they did not leverage projection-averaging to scale to higher dimensions and did not consider other distances between the CDFs. We take inspiration from distributionally robust control where linear projections of random

variables are used to define interpretable and easy-to-handle constraints [44]. *By providing the first (to our knowledge) proof-of-concept applications, through simple gradient-based algorithms, to distribution steering and ergodic control, we demonstrate that the introduced family of distances can be effectively used in control.*

The remainder of the chapter is organized as follows. In Section 3.2, we discuss existing work on distances and sliced variants. In Section 3.3 we provide a unification by introducing the proposed family of distances between distributions. In Section 3.4, we prove asymptotic results for estimators of the distance. In Section 3.5, we describe a practical algorithm to approximately compute the distance. In Section 3.6, we show the performance of this algorithm in two-sample tests. Finally, using the distance, in Section 3.8 we compute a control sequence for ergodic control, while in Section 3.7 we determine controllers for distribution steering. We conclude the chapter in Section 3.9.

3.2 *Prior work*

We discuss various projection-based strategies in turn and compare them against our proposed family of distances. Applications of such methods span generative modeling, clustering, and barycenter computation, where full multivariate optimal transport is infeasible [41], [45].

The sliced Wasserstein (SW) distance makes multivariate Wasserstein distances tractable by projecting the distributions onto one-dimensional subspaces and averaging their Wasserstein distances. The Wasserstein distance can be calculated in closed form in one dimension and involves the integral of the pointwise absolute difference of the CDFs [46]. Building on this idea, Kolouri et al. [42] proposed generalized sliced Wasserstein (GSW) distances using nonlinear projections, as well as max-sliced variants that optimize, rather than average, over projections. Projection-robust sliced Wasserstein distances that seek maximally discriminative subspaces have also ben suggested [47]. *The SW is included in our introduced family of distances.*

The Gromov–Wasserstein distance looks at the minimum expected absolute difference of distances between points generated by each distribution. The minimum is over the coupling of the distributions, setting the two distributions as the marginals. In the one-dimensional case, under assumptions, the distance can be computed by solving a quadratic assignment problem. By performing random one-dimensional projections of the two distributions and averaging their Gromov–Wasserstein distance we obtain the sliced variant [48]. The sliced Sinkhorn distance relies on the entropy-regularized optimal transport cost and can be computed by solving the dual of a convex program [37]. *Our distance does not require solving any inner optimization problem.*

Closely related to our proposed family of sliced distances are two statistics used for multivariate two-sample tests; the projection-averaged Cramer-von-Mises statistic and the projection-averaged Kolmogorov-Smirnov (KS) statistic. The former applies the discrepancy function $d(f,g) = \int_{-\infty}^{\infty}(f(x) - g(x))^2 dx$ to the CDFs of the random one-dimensional projections and is shown to have an efficient closed-form solution by Kim et al. [49] and by Li et al. [50]. The latter applies the discrepancy function $d(f,g) = \sup_x |f(x) - g(x)|$ to the CDFs of the random one-dimensional projections [39]. *The KS statistic can be used as the dissimilarity function in our proposed family of distances. The square root of the Cramer-von-Mises statistic can similarly be used as the dissimilarity function.*

The maximum mean discrepancy (MMD) compares two distributions through the largest difference in expectations over a prescribed class of functions $\mathcal{F}$,

$$\mathrm{MMD}_{\mathcal{F}}(P,Q) = \sup_{f\in\mathcal{F}} \left|\mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{Y\sim Q}[f(Y)]\right|. \tag{3.1}$$

When $\mathcal{F}$ is the unit ball of a reproducing kernel Hilbert space, the resulting discrepancy is equal to the distance between the kernel mean embeddings of the two distributions and can be evaluated using kernel expectations [51], [52], [53]. From a projection-based perspective, the functions f may be viewed as potentially nonlinear scalar projections of the random variables. MMD then selects the projection that maximizes the difference between the projected means, rather than comparing

the full one-dimensional projected distributions. With an appropriate kernel, this collection of nonlinear projections is sufficiently rich for the discrepancy to distinguish probability distributions [54]. *In contrast, our proposed family explicitly applies a univariate distributional dissimilarity to each projected distribution and then aggregates the resulting discrepancies over a chosen set of projections.*

3.3 The family of distances between distributions

Suppose $X : \Omega_X \to \mathbb{R}^n$ and $Y : \Omega_Y \to \mathbb{R}^n$ are two random variables, defined over potentially different probability spaces $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$ and $(\Omega_Y, \mathcal{F}_Y, \mathbb{P}_Y)$, respectively. We introduce a class of distances between the laws (or induced probability measures on $\mathbb{R}^n$) of X and Y that is based on linear one-dimensional projections of X and Y.

We first define the distance for the set $\mathcal{C}$ of univariate CDFs $f : \mathbb{R} \to [0, 1]$.

Definition 3.1. (**Distance in** $\mathcal{C}$). A function $d : \mathcal{C} \times \mathcal{C} \to \mathbb{R} \cup \{\infty\}$ is a distance for the set $\mathcal{C}$ if and only if it satisfies the following axioms:

1. *Non-negativity*: $d(f, g) \geq 0$,
2. *Symmetry*: $d(f, g) = d(g, f)$,
3. *Triangle inequality*: $d(f, g) \leq d(f, h) + d(h, g)$,
4. *Identity of indiscernibles*: $d(f, g) = 0 \Leftrightarrow f(x) = g(x)$ for all x.

Next, we define the one-dimensional linear projection of an n-dimensional random variable.

Definition 3.2. (**One-dimensional linear projection of a random variable**). Let $\tilde{q} \in \mathbb{R}^n$ be a vector. Then, $X_{\tilde{q}} := \tilde{q}^\top X$ is a linear one-dimensional projection of X.

By considering $X_{\tilde{q}}$ as a weighted sum of the coordinates of X, it is obvious that $X_{\tilde{q}}$ is a random variable over $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$, as a Borel function of X. Because $X_{\tilde{q}}$ is a one-dimensional random variable, it admits a CDF, $F_X^{\tilde{q}}(t) := \mathbb{P}\left(\tilde{q}^\top X \leq t\right)$.

We now define our measure of similarity between the laws (distributions) of X and Y.

Definition 3.3. (**Measure of similarity between distributions**). Suppose $\mathcal{P}_q := (\Omega_q, \mathcal{F}_q, \mathbb{P}_q)$ is a probability space and $q : \Omega_q \to \mathbb{S}^n$ is a random variable over $\mathcal{P}_q$, where $\mathbb{S}^n$ is the unit sphere in n dimensions. Suppose d is a distance in $\mathcal{C}$. Letting $\omega' \in \Omega_q$, we consider the random variable $d(F_X^{q(\omega')}, F_Y^{q(\omega')})$ over $\mathcal{P}_q$. We define the following measure of similarity between the laws of random variables X and Y, respectively:

$$\Delta(X, Y) := \mathbb{E}_q \left[d(F_X^q, F_Y^q) \right]. \tag{3.2}$$

Note that the random variable q takes values only on the unit sphere. This is a choice justified by the fact that $F_X^q(t) = F_X^{\frac{q}{\|q\|}}(t/\|q\|)$ for $q \neq 0$. Moreover, although presented in the context of $\mathbb{R}^n$, $\Delta(X, Y)$ is also applicable for distributions over matrices because matrices can be equivalently viewed as vectors (by stacking their columns or rows). It is important to note that $\Delta(X, Y)$ depends on the distribution of the random variable q on the unit sphere. A common choice is to use a uniformly distributed q, but, as we will see later on, this is not the only one. The uniform distribution is a good choice if we believe the coordinates of X and Y are similar in magnitude. If this is not the case, we should upweight the small coordinates of the random variables by using a non-uniform distribution for q (with larger values in the coordinates of q corresponding to the small coordinates of X and Y).

We now prove that $\Delta(X, Y)$ is a distance in the space of distributions.

Theorem 3.4. (***Distance property***). *$\Delta(X, Y)$, as given in (3.2), is a distance in the space of distributions.*

Proof. The non-negativity property of Δ follows from the non-negativity property of d. Similarly, the symmetry property of Δ follows from the symmetry property of d.

We now prove the triangle inequality. Consider an additional random variable $Z : \Omega_Z \to \mathbb{R}^n$ over $(\Omega_Z, \mathcal{F}_Z, \mathbb{P}_Z)$. Then:

$$\begin{aligned} \Delta(X, Y) &:= \mathbb{E}_q \left[d(F_X^q, F_Y^q) \right] \\ &\leq \mathbb{E}_q \left[d(F_X^q, F_Z^q) + d(F_Z^q, F_Y^q) \right] \\ &= \Delta(X, Z) + \Delta(Z, Y). \end{aligned} \tag{3.3}$$

Finally, we prove the identity of indiscernibles.

($\Rightarrow$). Suppose X and Y are equal in distribution: $X \overset{D}{=} Y$. Then, for all $\tilde{q} \in \mathbb{S}^n$, it holds that $F_X^{\tilde{q}} = F_Y^{\tilde{q}}$ pointwise. This implies that $d(F_X^{\tilde{q}}, F_Y^{\tilde{q}}) = 0$ for all $\tilde{q} \in \mathbb{S}^n$, by the zero identity of d. Therefore $\mathbb{E}_q\left[d(F_X^q, F_Y^q)\right] = 0$ and we obtain $\Delta(X, Y) = 0$.

($\Leftarrow$). Suppose $\Delta(X, Y) = 0$. Then $\mathbb{E}_q\left[d(F_X^q, F_Y^q)\right] = 0$. By Markov's inequality, for any $\alpha > 0$

$$\mathbb{P}_q\left(d(F_X^q, F_Y^q) \geq \alpha\right) \leq \frac{\mathbb{E}_q\left[d(F_X^q, F_Y^q)\right]}{\alpha} = 0. \tag{3.4}$$

Therefore $d(F_X^q, F_Y^q) = 0$ with probability 1, which implies that $F_X^q = F_Y^q$ pointwise with probability 1. Assuming the support of q is all $\mathbb{S}^n$, we obtain that for all $\tilde{q} \in \mathbb{S}^n$: $\tilde{q}^\top X \overset{D}{=} \tilde{q}^\top Y$, which implies that for all $\tilde{q} \in \mathbb{R}^n$: $\tilde{q}^\top X \overset{D}{=} \tilde{q}^\top Y$. By applying the Cramér–Wold theorem [55], it follows that $X \overset{D}{=} Y$. Intuitively, if the two random vectors X and Y agree on each projection, then they have the same distribution. □

Specific cases

Any function d that satisfies the requirements of Theorem 3.1 can be used in $\Delta(X, Y)$. Here, we outline a few choices.

- Suppose $d(f, g) = \sup_x |f(x) - g(x)|$. In this case, $\Delta(X, Y)$ is the expected KS distance between linear one-dimensional projections of X and Y. This choice of d can be particularly helpful, if we care about the maximum disagreement between the CDFs and not where the disagreement occurs.

- Suppose that $d(f, g) = \int_{-\infty}^{\infty} |f(x) - g(x)|dx$. Then, $\Delta(X, Y)$ captures the expected total difference between the CDFs of linear one-dimensional projections of X and Y. This is equivalent to the 1-Wasserstein distance on distributions. Note that we can generalize to $d(f, g) = \int_{-\infty}^{\infty} c(x)|f(x) - g(x)|dx$ for function c such that $c(x) > 0$ for all x. This can be helpful, if there are specific values of x for which we care about the agreement of the two CDFs.

- We can generalize the above point to $d(f, g) = \left(\int c(x)|f(x) - g(x)|^p dx\right)^{1/p}$ for $p \geq 1$. In other words, our family admits L^p distances between the CDFs

of the linear one-dimensional projections of X and Y.

3.4 A sample-based estimator

We define a sample-based estimator of $\Delta(X, Y)$. For simplicity we will assume that $F_Y^{\tilde{q}}$ is known for all fixed $\tilde{q}$. This can occur, for example, if we wish to compute the distance of the distribution of X to a known distribution; the distribution of Y.

Suppose we are given the samples (represented as random variables)

$$(q_i, X_{i,1}, \ldots, X_{i,N})_{i=1}^H$$

in a probability space with probability measure $\mathbb{Q} = \mathbb{P}_q \times (\times_{k=1}^N \mathbb{P}_X)$. The inner sample size N will be used to estimate CDFs, while the outer sample size H will be used to estimate the expectation in (3.2). The samples q_i are assumed independent and identically distributed to q. Similarly, the samples $X_{i,j}$ are assumed independent and identically distributed to X. Further, $(q_i)_{i=1}^H \perp (X_{i,1}, \ldots, X_{i,N})_{i=1}^H$, where $\perp$ denotes independence.

For each i, we can obtain an estimator of $F_X^{q_i,N}$ by calculating the empirical CDF of the samples $(X_{i,1}, \ldots, X_{i,N})$:

$$\hat{F}_X^{q_i,N}(t) =: \frac{1}{N} \sum_{j=1}^N \mathbb{1}\Big(q_i^\top X_{i,j} \leq t\Big). \tag{3.5}$$

We drop the dependence on N for simplicity below. We can now define an estimator for $\Delta(X, Y)$ as follows.

Definition 3.5. (**Estimator of** $\Delta(X, Y)$). Let $\hat{\Delta}(X, Y)$ be the following sample-based estimator of $\Delta(X, Y)$:

$$\hat{\Delta}(X, Y) =: \frac{1}{H} \sum_{i=1}^H d(\hat{F}_X^{q_i}, F_Y^{q_i}). \tag{3.6}$$

We study the asymptotic behavior of the estimator. First, we determine the asymptotic convergence of its mean and variance, and then provide an approximation for its asymptotic distribution.

Theorem 3.6. (***Asymptotic mean of estimator***). *Let $g : \mathcal{C} \to \mathbb{R}$. Suppose $d(\cdot, g) : \mathcal{C} \to \mathbb{R}$ is continuous with respect to the sup-norm in its first argument and bounded above uniformly by a constant M for all g. Then, the estimator $\hat{\Delta}(X, Y)$ is asymptotically unbiased, as $N \to \infty$.*

Proof. Obviously:

$$\mathbb{E}\left[\hat{\Delta}(X, Y)\right] = \mathbb{E}\left[d(\hat{F}_X^{q_i}, F_Y^{q_i})\right].$$

By the tower property of expectation

$$\mathbb{E}\left[d(\hat{F}_X^{q_i}, F_Y^{q_i})\right] = \mathbb{E}\left[\mathbb{E}\left[d(\hat{F}_X^{q_i}, F_Y^{q_i}) \mid q_i\right]\right], \tag{3.7}$$

and, in turn, by the independence assumptions, for any realization $\tilde{q}$ of q,

$$\mathbb{E}\left[d(\hat{F}_X^{q_i}, F_Y^{q_i}) \mid q_i = \tilde{q}\right] = \mathbb{E}_{\mathbb{P}}\left[d(\hat{F}_X^{\tilde{q}}, F_Y^{\tilde{q}})\right], \tag{3.8}$$

where $\mathbb{P} = \times_{i=1}^{N} \mathbb{P}_X$. Eq. (3.8) states that the conditional expectation of the distance for a given realization of q is equal to the expected distance for the given linear one-dimensional projection.

By the Glivenko—Cantelli theorem,

$$\mathbb{P}\left(\sup_x |\hat{F}_X^{\tilde{q}}(x) - F_X^{\tilde{q}}(x)| \to 0\right) = 1, \tag{3.9}$$

where the limit is for $N \to \infty$. The continuity assumption of d implies that

$$\begin{aligned} 1 &= \mathbb{P}\left(\sup_x |\hat{F}_X^{\tilde{q}}(x) - F_X^{\tilde{q}}(x)| \to 0\right) \\ &\leq \mathbb{P}\left(d(\hat{F}_X^{\tilde{q}}, F_Y^{\tilde{q}}) \to d(F_X^{\tilde{q}}, F_Y^{\tilde{q}})\right) \end{aligned} \tag{3.10}$$

and therefore $d(\hat{F}_X^{\tilde{q}}, F_Y^{\tilde{q}}) \xrightarrow{a.s.} d(F_X^{\tilde{q}}, F_Y^{\tilde{q}})$ as $N \to \infty$. Because d is bounded, we apply the dominated convergence theorem to get

$$\mathbb{E}_{\mathbb{P}}\left[d(\hat{F}_X^{\tilde{q}}, F_Y^{\tilde{q}})\right] \to d(F_X^{\tilde{q}}, F_Y^{\tilde{q}}) \text{ as } N \to \infty. \tag{3.11}$$

This result holds for every realization $\tilde{q}$.

The boundedness of d implies that $\mathbb{E}_{\mathbb{P}}\left[d(\hat{F}_X^{\tilde{q}}, F_Y^{\tilde{q}})\right]$ is also bounded by the same constant for any realization $\tilde{q}$. Therefore, the random variable $\mathbb{E}\left[d(\hat{F}_X^{q_i}, F_Y^{q_i}) \mid q_i\right]$ is bounded and, by applying the dominated convergence theorem to it, we get

$$\mathbb{E}\left[\hat{\Delta}(X,Y)\right] \to \mathbb{E}_q\left[d(F_X^q, F_Y^q)\right] = \Delta(X,Y) \tag{3.12}$$

as $N \to \infty$. □

Various choices of d satisfy the continuity condition required by Theorem 3.6. For example, suppose $d(f,g) = \max_x |f(x) - g(x)|$. Consider the sequence of functions $f_n(x)$ such that $\max_x |f_n(x) - f(x)| \to 0$. Then:

$$d(f_n, g) \leq \max_x |f_n(x) - f(x)| + \max_x |f(x) - g(x)|,$$
$$d(f, g) \leq \max_x |f_n(x) - f(x)| + \max_x |f_n(x) - g(x)|,$$

and we obtain

$$\lim_{n\to\infty} d(f_n, g) \leq d(f,g), \quad \lim_{n\to\infty} d(f_n, g) \geq d(f,g),$$

i.e., $\lim_{n\to\infty} d(f_n, g) = d(f,g)$. The same result holds for $d(f,g) = \int_{-\infty}^{\infty} c(x)|f(x) - g(x)|dx$, where c is a probability density function (PDF). To see this, the reader can follow the steps outlined above.

Furthermore, because CDFs are constrained between 0 and 1, we see that both $d(f,g) = \max_x |f(x) - g(x)|$ and $d(f,g) = \int_{-\infty}^{\infty} c(x)|f(x) - g(x)|dx$ satisfy $d(f,g) \leq 1$, i.e., are bounded.

Corollary 3.7. (***Asymptotic variance of estimator***). *Let $g : \mathcal{C} \to \mathbb{R}$. Suppose $d(\cdot, g) : \mathcal{C} \to \mathbb{R}$ is continuous with respect to the sup-norm in its first argument and bounded above uniformly by a constant M for all g. Then,*

$$\text{var}\left[\hat{\Delta}(X,Y)\right] \to \frac{1}{H}\text{var}_q\left[d(F_X^q, F_Y^q)\right] \text{ as } N \to \infty. \tag{3.13}$$

Proof. Obviously

$$\text{var}\left[\hat{\Delta}(X,Y)\right] = \frac{1}{H}\text{var}\left[d(\hat{F}_X^{q_i}, F_Y^{q_i})\right] \tag{3.14}$$

and

$$\begin{aligned}&\text{var}\left[d(\hat{F}_X^{q_i}, F_Y^{q_i})\right] = \\ &\mathbb{E}\left[d^2(\hat{F}_X^{q_i}, F_Y^{q_i})\right] - \mathbb{E}\left[d(\hat{F}_X^{q_i}, F_Y^{q_i})\right]^2.\end{aligned} \tag{3.15}$$

We study the asymptotic behavior of the two terms on the right-hand side in turn. By a similar argument as in Theorem 3.6, we obtain

$$\mathbb{P}\left(d^2(\hat{F}_X^{\tilde{q}}, F_Y^{\tilde{q}}) \to d^2(F_X^{\tilde{q}}, F_Y^{\tilde{q}})\right) = 1 \tag{3.16}$$

for all realizations $\tilde{q}$. Because d is bounded, by applying the dominated convergence theorem, we get

$$\mathbb{E}_{\mathbb{P}}\left[d^2(\hat{F}_X^{\tilde{q}}, F_Y^{\tilde{q}})\right] \to d^2(F_X^{\tilde{q}}, F_Y^{\tilde{q}}) \text{ as } N \to \infty. \tag{3.17}$$

A second application of the dominated convergence theorem, this time for the sequence of random variables $\mathbb{E}\left[d^2(\hat{F}_X^{q_i}, F_Y^{q_i}) \mid q_i\right]$, gives

$$\mathbb{E}\left[d^2(\hat{F}_X^{q_i}, F_Y^{q_i})\right] \to \mathbb{E}_q\left[d^2(F_X^{q}, F_Y^{q})\right] \text{ as } N \to \infty. \tag{3.18}$$

By Theorem 3.6, we directly get

$$\mathbb{E}\left[d(\hat{F}_X^{q_i}, F_Y^{q_i})\right]^2 \to \Delta^2(X,Y) \text{ as } N \to \infty. \tag{3.19}$$

By combining (3.18) and (3.19), we get the result. □

Theorem 3.6 and Theorem 3.7 provide the asymptotic mean and variance of $\hat{\Delta}(X,Y)$, respectively. We can also determine the asymptotic distribution of $\hat{\Delta}(X,Y)$.

Theorem 3.8. (***Asymptotic distribution of estimator***). *Let $g : \mathcal{C} \to \mathbb{R}$. Suppose $d(\cdot, g) : \mathcal{C} \to \mathbb{R}$ is continuous with respect to the sup-norm in its first argument and*

bounded above uniformly by a constant M for all g. Then,

$$\sqrt{H}\left(\hat{\Delta}(X,Y) - \Delta(X,Y)\right) \xrightarrow{D} \mathcal{N}\left(0, \mathrm{var}_q\left[d(F_X^q, F_Y^q)\right]\right), \tag{3.20}$$

as $H, N \to \infty$.

Proof. Suppose N_H is a strictly increasing function of H. Consider the triangular array of random variables

$$\begin{array}{cccccc} Z_1^{(1)} & & & & & \\ Z_2^{(1)} & Z_s^{(2)} & & & & \\ \vdots & \vdots & \ddots & & & , \\ Z_H^{(1)} & Z_H^{(2)} & \dots & Z_H^{(H)} & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \end{array}$$

where

$$\begin{aligned} &\left(Z_H^{(i)}\right)_{i=1}^{H} =: \\ &\left(d(\hat{F}_X^{q_i,N_H}, F_Y^{q_i}) - \mathbb{E}\left[d(\hat{F}_X^{q_i,N_H}, F_Y^{q_i})\right]\right)_{i=1}^{H}, \end{aligned} \tag{3.21}$$

and each $\hat{F}_X^{q_i,N_H}$ depends on $(X_{i,k})_{k=1}^{N_H}$. Obviously

$$\mathbb{E}\left[Z_H^{(i)}\right] = 0, \quad \mathrm{var}\left[Z_H^{(i)}\right] = \mathrm{var}\left[d(\hat{F}_X^{q_1,N_H}, F_Y^{q_1})\right].$$

By Theorem 3.7, $\mathrm{var}\left[Z_H^{(i)}\right] \to \mathrm{var}_q\left[d(F_X^q, F_Y^q)\right]$ as $H \to \infty$. We define

$$\bar{Z}^{(H)} := \frac{1}{H}\sum_{i=1}^{H} Z_i^{(H)}.$$

Because the distance d is bounded, we can easily confirm that the Lindeberg condition

$$\frac{1}{H}\sum_{i=1}^{H} \mathbb{E}\left[\left(Z_i^{(H)}\right)^2 \mathbb{1}\left(|Z_i^{(H)}| \geq \epsilon\sqrt{H}\right)\right] \to 0, \text{ as } H \to \infty \tag{3.22}$$

holds for all $\epsilon > 0$. By applying the Lindeberg central limit theorem [56], we obtain

$$\sqrt{H}\bar{Z}^{(H)} \xrightarrow{D} \mathcal{N}\left(0, \mathrm{var}_q\left[d(F_X^q, F_Y^q)\right]\right), \text{ as } H \to \infty. \tag{3.23}$$

This implies the desired result. □

Theorem 3.8 proves the intuitive fact that an estimator based on empirical averages is asymptotically normally distributed. It also allows us to better understand the behavior of $\hat{\Delta}(X, Y)$. Specifically, suppose $\Delta(X, Y) = \alpha$, where $\alpha \in [0, 1]$. By approximating $d(F_X^q, F_Y^q)$ as a Bernoulli random variable with probability of success α, we get that $\mathrm{var}_q\left[d(F_X^q, F_Y^q)\right] = \alpha(1-\alpha)$. Therefore, the asymptotic signal-to-noise ratio for the estimator is $\sqrt{H\alpha/(1-\alpha)}$. This is an increasing function of α. This, in turn, means that it is easier to conclude that two distributions are different than to conclude that two distributions are similar. Further, by increasing the number of samples H, we can reduce the standard error as desired.

3.5 *A practical algorithm*

We suggest a practical way to approximately compute $\Delta(X, Y)$ for a specific choice of the function d. For this choice of d we argue that $\Delta(X, Y)$ is easily computable and interpretable. We will assume

$$d(F_X^q, F_Y^q) =: \int_{-\infty}^{\infty} c(x \mid q)|F_X^q(x) - F_Y^q(x)|dx, \tag{3.24}$$

where $c(x \mid q)$ is a PDF. We can write

$$\begin{aligned} &d(F_X^q, F_Y^q) \\ &= \int_{-\infty}^{\infty} c(x \mid q)|\mathbb{P}\left(q^\top X \geq x\right) - \mathbb{P}\left(q^\top Y \geq x\right)|dx \end{aligned} \tag{3.25}$$

and interpret d as the expected (with respect to PDF $c(\cdot|q)$) absolute difference in probability content of X and Y on half-spaces with normal vector q. In this case, we

can write

$$\Delta(X,Y) = \int_{\mathbb{S}^n}\int_{-\infty}^{\infty} p_q(\tilde{q})c(x \mid \tilde{q})|\mathbb{P}\left(\tilde{q}^\top X \geq x\right) - \mathbb{P}\left(\tilde{q}^\top Y \geq x\right)|dxd\tilde{q}, \tag{3.26}$$

where p_q is a PDF over $\mathbb{S}^n$. This distance shares similarities with the sliced Wasserstein distances [42] and the continuous ranked probability score [40].

Eq. (3.26) shows that $\Delta(X,Y)$ effectively samples random half-spaces and computes the difference in probability within each half-space for the two random variables X and Y. Therefore $\Delta(X,Y)$ can be interpreted as the expected difference in probability content within a half-space for the two distributions.

Because $|\mathbb{P}\left(\tilde{q}^\top X \geq x\right) - \mathbb{P}\left(\tilde{q}^\top Y \geq x\right)| \in [0,1]$, we notice that $\Delta(X,Y)$ is constrained between 0 and 1. Although, we will not discuss it further, this makes $\Delta(X,Y)$ ideal as a reward for sequential decision-making algorithms like Monte Carlo tree search (MCTS), where scaling the cost can significantly improve performance [57], [58], [59].

In various applications one of the random variables, X or Y, is represented as a set of samples, while the PDF of the other is given explicitly. This is common, for example, in distribution steering, where the target distribution is given as a PDF, while the current state is given as a set of samples. Suppose that the PDF of Y is given as a Gaussian mixture model (GMM). The GMM is a universal approximator, in the sense that any smooth density can be approximated with arbitrary error by a GMM [60]. We show that $\mathbb{P}\left(q^\top Y + b \geq 0\right)$ can be computed in closed form. This holds because

$$\mathbb{E}_{y\sim\mathcal{N}(\bar{y},\Sigma)}\mathbb{1}\left(q^T y + b \geq 0\right) = \mathbb{P}\left(r \geq 0\right), \tag{3.27}$$

where $r \sim \mathcal{N}(q^T\bar{y} + b, q^T\Sigma q)$ is a one-dimensional Gaussian. The last probability can be easily computed. Therefore, if Y follows a GMM density μ_Y, we can compute $\mathbb{E}_{y\sim\mu_Y}\mathbb{1}\left(q^T y + b \geq 0\right)$ as follows: we weight the probabilities of the form (3.27) for each component of the GMM by their corresponding weights and add the results.

Assuming X is given as a set of samples $\{x^i\}_{i=1}^N$ and the PDF of Y, μ_Y, is known explicitly, we show how to approximate $\Delta(X,Y)$ in Algorithm 3.1 with complexity

$\mathcal{O}\left(HN(n + n_{\text{values}})\right)$. We denote the output of Algorithm 3.1 as $D_{H,n_{\text{values}}}(X, Y)$.

Algorithm 3.1 Approximation of $\Delta(X, Y)$ with d as in (3.24)

1: **Inputs:** density μ_Y of Y, samples $\{x_i\}_{i=1}^N$ of X
2: **Parameters:** H, n_{values},
3: $\Delta \leftarrow 0$
4: **for** $k = 1, \ldots, H$
5: get a half-space $q_k \in \mathbb{S}^n$ according to the density $p_q(\cdot)$
6: **for** $j = 1, \ldots, n_{\text{values}}$
7: get a value $b_j \in \mathbb{R}$ according to the density $c(\cdot \mid q_k)$
8: $p_{\text{samples}} \leftarrow \frac{1}{N} \sum_{i=1}^N \mathbb{1}\left(q_k^\top x_i \geq b_j\right)$
9: $p_{\mu_Y} \leftarrow \mathbb{E}_{y \sim \mu_Y} \mathbb{1}\left(q_k^\top y \geq b_j\right)$
10: $\Delta \leftarrow \Delta + \frac{1}{H n_{\text{values}}} |p_{\text{samples}} - p_{\mu_Y}|$
11: **return** Δ

An important aspect in Algorithm 3.1 is the procedure for obtaining the vectors q_k and the values b_j. The approximate value for $\Delta(X, Y)$, $D_{H,n_{\text{values}}}(X, Y)$, depends on this choice. Although Algorithm 3.1 works for arbitrary densities $p_q(\cdot)$ and $c(\cdot|q)$, uniformly sampling the directions q_k from the unit sphere and picking the b_j as the quantiles of the sample set $\{q_k^\top x_i\}_{i=1}^N$ worked well in practice.

The modifications necessary to Algorithm 3.1 in the case that Y is represented as a set of samples or the PDF of X is given explicitly are obvious.

3.5.1 *Differentiability considerations*

In many applications, it is important for the estimate $D_{H,n_{\text{values}}}(X, Y)$ to be differentiable with respect to the samples $\{x_i\}_{i=1}^N$. For example, in distribution steering we want to find a controller that steers the distribution of the samples $\{x_i\}_{i=1}^N$ towards a target distribution μ_Y. The estimate provided by Algorithm 3.1 is not differentiable, because the indicator function in line 8 is not differentiable. We can obtain a differentiable modification of the estimate by replacing $\mathbb{1}(q_k^\top x_i \geq b_j)$ with $\sigma\left(V(q_k^\top x_i - b_j)\right)$, where σ is the sigmoid function and V is a large scalar. We pick $V = 100$, which makes the resulting function steep near the origin. Below, when taking the gradient

of $D_{H,n_{\text{values}}}(X, Y)$ we assume the sigmoid approximation is used. The gradient of $D_{H,n_{\text{values}}}(X, Y)$ is a stochastic gradient of $\Delta(X, Y)$.

3.6 Two-sample tests

Any distance between two distributions (or, empirically, between two sets of samples) should satisfy two important properties.

- It should be able to tell whether two sets of samples are produced by the same distribution.
- It should be able to tell whether two sets of samples are produced by different distributions.

This is the purpose of a two-sample test, which is a statistical test that determines whether two sets of samples are produced by the same distribution or not.

The set-up is as follows. We consider two sets of samples $\{X_i\}_{i=1}^N$ and $\{Y_i\}_{i=1}^N$ in $\mathbb{R}^n$ produced by distributions μ_X and μ_Y, respectively, and the two cases for μ_X and μ_Y:

$$\mathrm{H}_0 : \mu_X = \mu_Y, \quad \mathrm{H}_\mathrm{A} : \mu_X \neq \mu_Y. \tag{3.28}$$

We then look at the distribution of the distance between $\{X_i\}_{i=1}^N$ and $\{Y_i\}_{i=1}^N$ in each case. If the distributions of the distance in the two cases are easily discernible, then we can conclude that the distance is effective in distinguishing between the two cases.

An important consideration is the need to decide upon a mechanism that takes as input the value of the distance and outputs a decision whether the samples are produced by the same distribution or not (accept or reject the null hypothesis H_0, respectively). This is typically done by selecting a threshold value for the distance. Selecting the threshold value is a trade-off between the probability of false positives (rejecting H_0 when it is true) and the probability of false negatives (accepting H_0 when it is false). It is a challenging task, especially when we only have access to the samples $\{X_i\}_{i=1}^N$ and $\{Y_i\}_{i=1}^N$. A simple approach is to select the threshold value

as a percentile of the distribution of the distance between subsets of samples from $\{X_i\}_{i=1}^N$. The intuition is that for these subsets of samples, the null hypothesis H_0 is true, and therefore the distribution of the distance between these subsets of samples is a good approximation of the distribution of the distance in the null case. The distance should typically be larger than the threshold value in the alternative case.

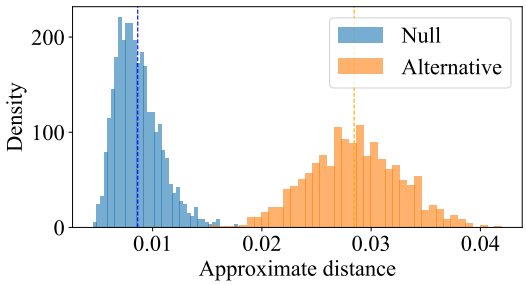
In this section, we examine the performance of the distance estimate $D_{H,n_{\text{values}}}(X, Y)$ by Algorithm 3.1 in this context. First, we look at the case involving only multivariate Gaussian distributions, and then we consider a challenging case of non-Gaussian distributions.

3.6.1 *The Gaussian case*

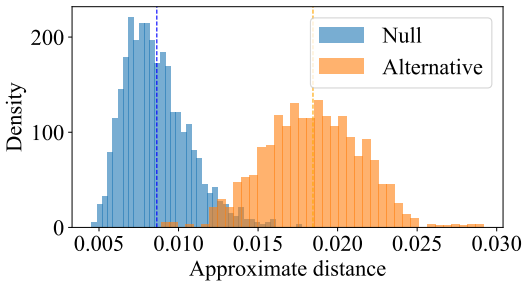
We look at the empirical distribution of $D_{H,n_{\text{values}}}(X, Y)$ in the null H_0 and alternative H_A cases for $n = 2$ and $n = 50$. The randomness in $D_{H,n_{\text{values}}}(X, Y)$ comes from the drawn samples, $\{X_i\}_{i=1}^N$ and $\{Y_i\}_{i=1}^N$, and the half-spaces in Algorithm 3.1. We set N= 1000, H= 300 and n_{values}= 100. In the null case H_0 the samples are produced from μ_X and μ_Y that are $\mathcal{N}(0, I)$. For each n, we consider two alternative cases: $(\mu_X = \mathcal{N}(0, I), \mu_Y = \mathcal{N}(0.2, I))$ and $(\mu_X = \mathcal{N}(0, I), \mu_Y = \mathcal{N}(0, 1.3I))$. The results are included in Figure 3.2. The distributions of $D_{H,n_{\text{values}}}(X, Y)$ in the null and alternative case are easily discernible.

3.6.2 *A non-Gaussian case*

We compare the behavior of the proposed sliced distance in Algorithm 3.1 with several commonly used discrepancies between multivariate distributions, as a function of the dimension n. The objective of this experiment is to evaluate how effectively each discrepancy distinguishes two distributions that have identical first and second moments but differ in their higher-order structure. This setting is particularly relevant for distributional comparison because methods that primarily capture differences in location or covariance cannot distinguish the two populations in this case.

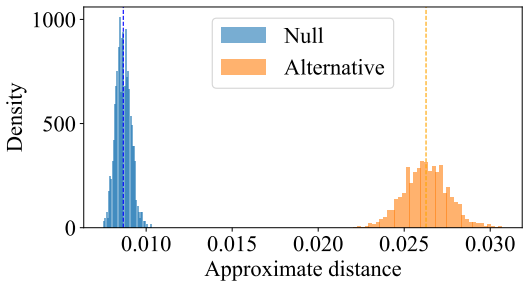


Alternative $(\mu_X = \mathcal{N}(0, I),\ \mu_Y = \mathcal{N}(0.2, I))$.

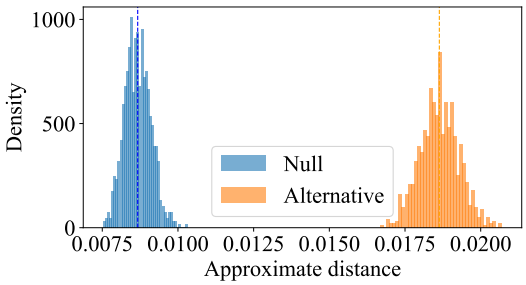


Alternative $(\mu_X = \mathcal{N}(0, I),\ \mu_Y = \mathcal{N}(0, 1.3I))$.

Dimensionality $n = 2$.



Alternative $(\mu_X = \mathcal{N}(0, I),\ \mu_Y = \mathcal{N}(0.2, I))$.



Alternative $(\mu_X = \mathcal{N}(0, I),\ \mu_Y = \mathcal{N}(0, 1.3I))$.

Dimensionality $n = 50$.

Figure 3.2: The empirical distribution of $D_{300,100}(X, Y)$ for the null case and different alternative cases in the two-sample test for Gaussian distributions. The randomness comes from the drawn samples for X and Y and the selection of half-spaces in Algorithm 3.1. In all experiments, the null case is $\mu_X = \mu_Y = \mathcal{N}(0, I)$. Each experiment corresponds to a particular choice of the dimensionality n and the alternative case. The distribution of $D_{300,100}(X, Y)$ is shown in blue for the null case and in orange for the alternative. The mean values are denoted by vertical dotted lines.

For each dimension n, we consider a null and an alternative experiment. Under the null experiment, both sample sets are generated independently from the same standard multivariate Gaussian distribution,

$$X \sim \mathcal{N}(0, I_n), \qquad Y \sim \mathcal{N}(0, I_n). \tag{3.29}$$

Under the alternative hypothesis, the first sample set remains Gaussian, while the second is generated from a product Laplace distribution,

$$X \sim \mathcal{N}(0, I_n), \qquad Y_j \overset{\text{iid}}{\sim} \text{Laplace}\left(0, \frac{1}{\sqrt{2}}\right), \quad j = 1, \ldots, n. \tag{3.30}$$

A scalar Laplace random variable with scale parameter b has variance $2b^2$. The choice $b = 1/\sqrt{2}$ therefore ensures that each coordinate of Y has unit variance. Consequently, the Gaussian and Laplace distributions in the alternative experiment have the same mean vector and covariance matrix,

$$\mathbb{E}[X] = \mathbb{E}[Y] = 0, \qquad \text{Cov}(X) = \text{Cov}(Y) = I_n. \tag{3.31}$$

The two distributions differ only through higher-order features such as tail behavior, kurtosis, and the shape of their level sets. Their PDFs are illustrated in Figure 3.3. The experiment therefore measures the ability of each discrepancy to detect differences beyond the first two moments.

For every dimension, we generate two independent empirical sample sets of size $N = 1000$. The experiment is repeated over multiple Monte Carlo trials. In each trial, we compute the proposed sliced CDF distance, a kernel maximum mean discrepancy, an empirical Wasserstein distance, and a kernel-density-based estimate of the Kullback–Leibler divergence. For each distance/discrepancy, we report the mean and standard deviation across trials under both the null and alternative distributions.

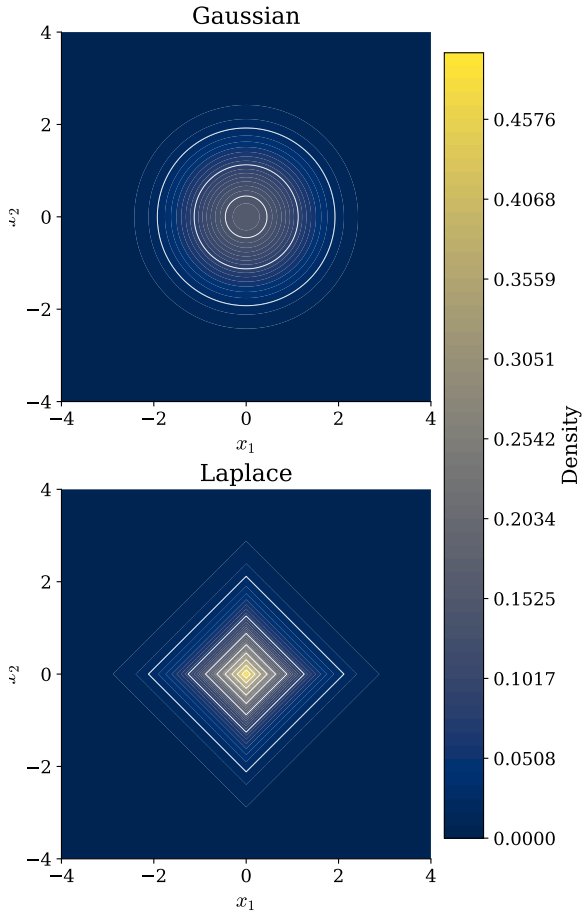


Figure 3.3: Comparison of the two distributions in (3.30). The distributions differ only in higher-order characteristics. We denote the PDFs and the level sets.

Sliced CDF distance. For the sliced comparison, we independently sample L directions

$$q_\ell \sim \mathrm{Unif}\left(\mathbb{S}^{n-1}\right), \qquad \ell = 1, \dots, L, \tag{3.32}$$

by drawing standard Gaussian vectors and normalizing them to unit Euclidean norm. We use $L = 600$. For each direction, the samples are projected onto the real line,

$$x_i^{(\ell)} = q_\ell^\top X_i, \qquad y_j^{(\ell)} = q_\ell^\top Y_j. \tag{3.33}$$

We then compute the L^1 distance between the corresponding empirical cumulative distribution functions $\widehat{F}_{X,\ell}(t)$ and $\widehat{F}_{Y,\ell}(t)$,

$$d_{\mathrm{CDF}}\left(q_\ell^\top X, q_\ell^\top Y\right) = \int_{-\infty}^{\infty} \left|\widehat{F}_{X,\ell}(t) - \widehat{F}_{Y,\ell}(t)\right| \, dt. \tag{3.34}$$

For empirical CDFs on the real line, this integral can be evaluated exactly by partitioning the real line at the pooled projected observations. It is also equal to the empirical one-dimensional 1-Wasserstein distance. The final sliced discrepancy is the average over sampled directions,

$$\widehat{d}_{\mathrm{sliced}}(X, Y) = \frac{1}{L} \sum_{\ell=1}^{L} d_{\mathrm{CDF}}\left(q_\ell^\top X, q_\ell^\top Y\right). \tag{3.35}$$

Thus, no density estimation or inner optimization problem is required. The only approximation is the finite Monte Carlo average over projection directions.

Maximum mean discrepancy. We also compute the maximum mean discrepancy using a Gaussian radial basis function kernel,

$$k_\sigma(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2}\right). \tag{3.36}$$

Given empirical samples $\{X_i\}_{i=1}^N$ and $\{Y_j\}_{j=1}^N$, we use the biased, nonnegative estimator

$$\widehat{\mathrm{MMD}}^2(X,Y) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N} k_\sigma(X_i, X_{i'}) + \frac{1}{N^2}\sum_{j=1}^{N}\sum_{j'=1}^{N} k_\sigma(Y_j, Y_{j'}) - \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k_\sigma(X_i, Y_j). \tag{3.37}$$

We report $\widehat{\mathrm{MMD}}$, rather than its square. The biased estimator includes the diagonal kernel terms and is equal to the norm of the difference between the two empirical kernel mean embeddings. It is therefore nonnegative up to numerical precision.

The kernel bandwidth is selected separately in each trial using the median heuristic. In particular, σ is set equal to the median of the nonzero pairwise Euclidean distances in the pooled sample. We use a pooled sample of 500 datapoints. To limit the cost of the bandwidth calculation, the median is estimated using at most a fixed number of pooled observations when the total sample size is large. The full sample sets are nevertheless used in the MMD estimator itself. This bandwidth choice adapts the kernel scale to the typical distance between observations, although its dependence on dimension should be kept in mind when interpreting the results.

Empirical Wasserstein distance. As a multivariate optimal-transport benchmark, we compute the empirical 2-Wasserstein distance. For two equally weighted empirical distributions with the same number of observations, this takes the form

$$\widehat{W}_2^2(X,Y) = \min_{\pi \in S_n} \frac{1}{n}\sum_{i=1}^{n} \left\| X_i - Y_{\pi(i)} \right\|_2^2, \tag{3.38}$$

where S_n denotes the set of permutations of $\{1,\dots,n\}$. The optimal permutation is obtained by solving a linear assignment problem with pairwise squared Euclidean costs. We report the square root of the optimal average cost,

$$\widehat{W}_2(X,Y) = \left(\widehat{W}_2^2(X,Y)\right)^{1/2}. \tag{3.39}$$

We solve the assignment problem using the Hungarian algorithm [61], because the two sample sets are of equal size.

KDE-based Kullback–Leibler divergence. Finally, we construct a plug-in estimate of the directed Kullback–Leibler divergence

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{X\sim P}\left[\log p(X) - \log q(X)\right]. \tag{3.40}$$

We fit Gaussian kernel density estimates $\widehat{p}$ and $\widehat{q}$ to the two sample sets, $\{X_i\}_{i=1}^{N}$ and $\{Y_i\}_{i=1}^{N}$ respectively, and evaluate

$$\widehat{D}_{\mathrm{KL}}(P\|Q) = \frac{1}{N_{\mathrm{eval}}} \sum_{i=1}^{N_{\mathrm{eval}}} \left[\log \widehat{p}(\widetilde{X}_i) - \log \widehat{q}(\widetilde{X}_i)\right], \tag{3.41}$$

where the evaluation observations $\widetilde{X}_i$ are drawn independently from P. P corresponds to the standard Gaussian distribution in this section. Using an independent evaluation sample avoids evaluating the fitted density exclusively at its own training observations. The bandwidths are selected according to the default rule used by the Gaussian kernel density estimator.

This estimator is included to illustrate the difficulty of direct multivariate density estimation. As dimension increases, the kernel density estimates become increasingly data intensive, and the resulting KL estimates may exhibit high variance, bias, or numerical instability. Failed density-estimation trials are recorded as missing values and excluded when computing the Monte Carlo summary statistics.

Dimension sweep and evaluation. The dimension n is varied over a prescribed grid. For each value of n, the null and alternative experiments are repeated using independent random samples. Let $d_{\mathrm{null}}^{(r)}(n)$ and $d_{\mathrm{alt}}^{(r)}(n)$ denote the value of a given discrepancy in Monte Carlo trial r. We summarize its behavior using

$$\overline{d}_{\mathrm{null}}(n) = \frac{1}{R}\sum_{r=1}^{R} d_{\mathrm{null}}^{(r)}(n), \qquad \overline{d}_{\mathrm{alt}}(n) = \frac{1}{R}\sum_{r=1}^{R} d_{\mathrm{alt}}^{(r)}(n), \tag{3.42}$$

together with the corresponding sample standard deviations.

A useful discrepancy should remain small under the null while producing a visibly larger value under the alternative. Accordingly, the separation between the null and alternative curves measures the ability of the method to identify the difference between Gaussian and Laplace distributions with matched means and covariances. The width of the standard-deviation bands reflects the finite-sample variability of each estimator. This comparison does not place the four discrepancies on a common numerical scale; their absolute magnitudes are not directly comparable. Instead, the relevant comparison is how the null–alternative separation and estimator variability evolve with dimension for each method.

The experiment also highlights the different computational mechanisms used by the methods. The sliced CDF distance reduces the multivariate problem to exact one-dimensional empirical comparisons and averages over random linear projections. MMD compares empirical means after an implicit nonlinear feature transformation induced by the kernel. The empirical Wasserstein distance solves a multivariate optimal assignment problem, while the KL estimator requires direct multivariate density estimation. Their behavior as dimension increases therefore reflects both the geometry of the underlying discrepancy and the statistical or computational approximation used to evaluate it.

Results are shown in Figure 3.4. Overall, the sliced distance and the MMD exhibity the best performance across dimensions, although performance degrades as the dimension increases. The fact that the sliced distance and the MMD are on the same order of magnitude is not surprising: the sliced distance is bounded in $[0, 1]$ because it averages distances between CDFs, while the MMD depends on the kernel evaluations where we have set the badnwidth using the median heuristic. The empirical Wasserstein distance and the KDE-based KL divergence are less effective in distinguishing the two distributions, particularly in higher dimensions. The Wasserstein distance is more sensitive to the curse of dimensionality, while the KDE-based KL divergence suffers from the challenges of multivariate density estimation. The fact that the KDE-based KL divergence is negative is because we use fitted densities and evaluate the KL divergence on an independent sample. This

is a known issue with plug-in estimators of the KL divergence.

Sensitivity to the number of projections and datapoints

We next examine how the empirical sliced distance depends on the two main computational parameters of the estimator: the number of sampled half-spaces and the number of available datapoints. Throughout this experiment, the dimension is fixed at $n = 50$. This is a challenging setting as observed in Figure 3.4.

The first experiment studies the dependence of the sliced distance on the number of random projection directions. The sample size is fixed at $N = 1000$, while the number of half-spaces is varied over a logarithmically spaced grid.

The second experiment studies the dependence on the sample size N. The number of random half-spaces is fixed at 600, while the number of observations is varied over a logarithmically spaced grid.

For each parameter configuration, the experiment is repeated independently, and the solid curves in Figure 3.5 report the Monte Carlo mean of the sliced discrepancy. The shaded regions represent one standard deviation above and below the corresponding mean. The randomness in each trial comes from the drawn samples and the random half-spaces. We observe that the ability to distinguish the null and alternative cases depends mostly on the number of datapoints and not on the number of half-spaces.

3.7 *Distribution steering*

In distribution steering, we consider a distribution for the state of a dynamical system at any time t and the goal is to match the state distribution with a target distribution as $t \to \infty$. We suppose the dynamical system

$$x_{t+1} = f(x_t, u_t), \tag{3.43}$$

where $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^m$, and $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$. We suppose that the state at the initial time $t = 0$ follows a distribution μ_{start}, given as a set of samples $\{x_i^{(0)}\}_{i=1}^N$.

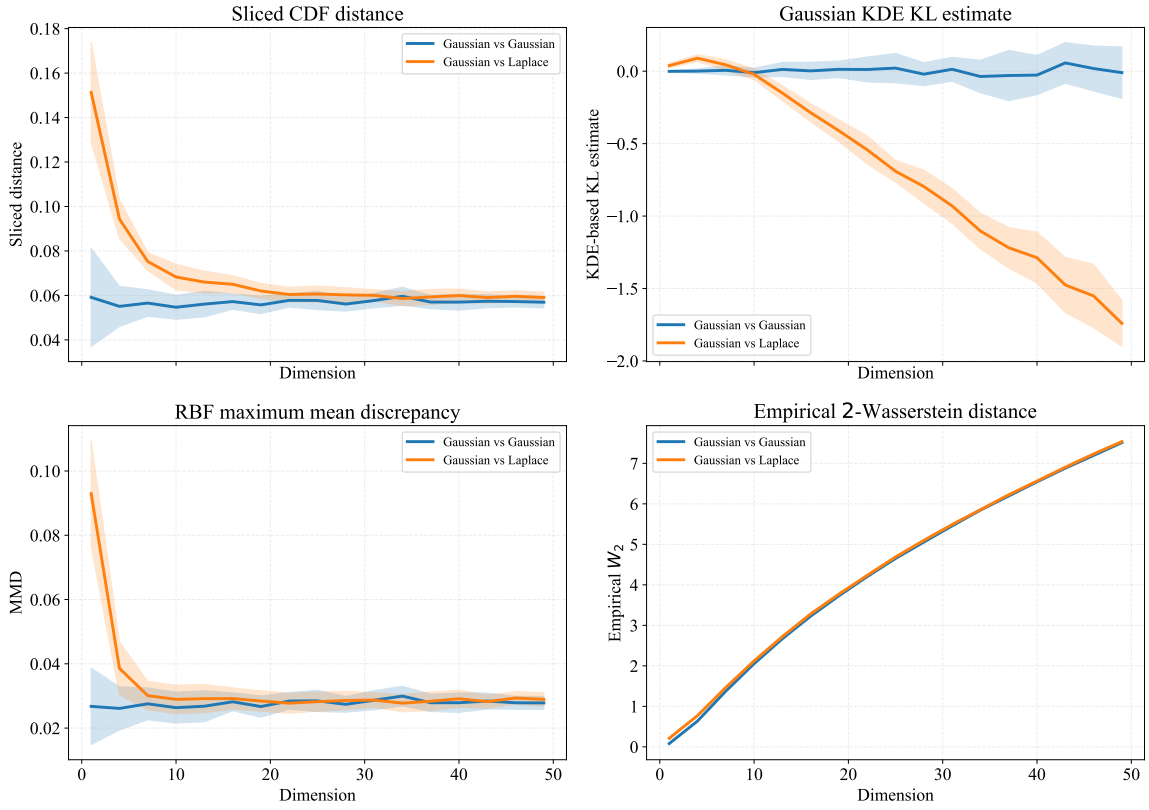


Figure 3.4: Comparison of distributional discrepancies as the dimension of the random vectors increases. For each dimension, two independent samples are generated under the null from $\mathcal{N}(0, I_n)$ and under the alternative from $\mathcal{N}(0, I_n)$ and an independent product Laplace distribution with the same mean and covariance. The panels report the sliced CDF distance, the KDE-based estimate of the Kullback–Leibler divergence, the Gaussian-kernel maximum mean discrepancy, and the empirical 2-Wasserstein distance. Solid lines show the mean over Monte Carlo trials, and the shaded regions indicate one sample standard deviation.

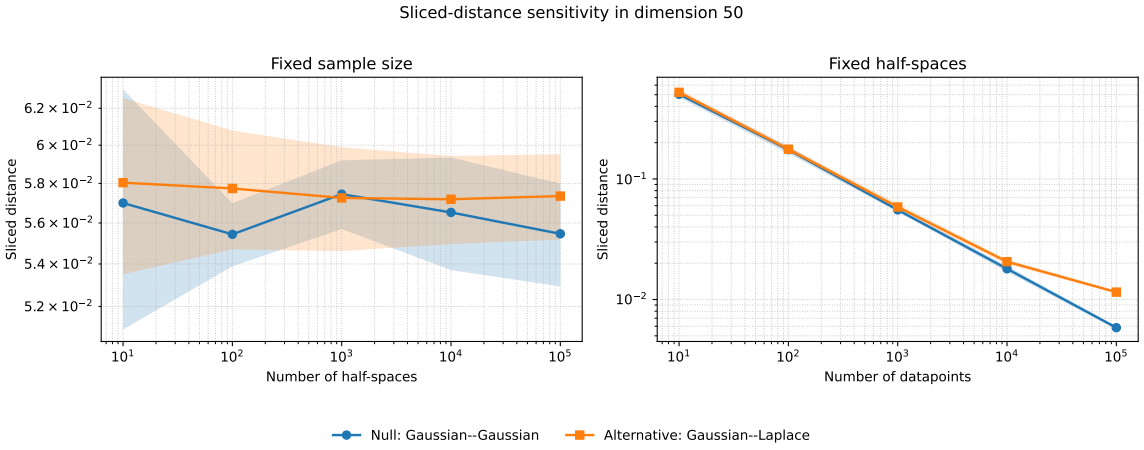


Figure 3.5: Sensitivity of the sliced discrepancy in dimension $n = 50$. The null experiment compares two independent samples from $\mathcal{N}(0, I_n)$, while the alternative compares a sample from $\mathcal{N}(0, I_n)$ with a sample having independent standardized Laplace coordinates. The two distributions under the alternative have the same mean and covariance. Left: the sample size is fixed at $N = 1000$, and the number of sampled half-spaces is varied. Right: the number of half-spaces is fixed at 600, and the number of datapoints is varied. Solid curves show the Monte Carlo mean, and shaded regions show one standard deviation. Both axes are displayed on logarithmic scales.

The distribution of the state at time t is represented by the set of samples $\{x_i^{(t)}\}_{i=1}^N$, in general.

Our task is to design a closed-loop controller that steers the distribution of the state at t towards the distribution μ_Y, as $t \to \infty$. We choose the controller family $u_t = K_t x_t + b_t$, with parameters (K_t, b_t). This is a common choice in the literature [30], [32]. Here μ_Y is the target distribution, which we assume is given as a PDF.

Our simple method for computing the controller at t, given the state distribution at t, $\{x_i^{(t)}\}_{i=1}^N$, is as follows. Suppose we start with the controller $\pi_t = (K_t, b_t)$. Then, the samples at the next time, $t+1$, will be $\{x_i^{(t+1)}\}_{i=1}^N$, where

$$x_i^{(t+1)} = f\left(x_i^{(t)}, \pi_t(x_i^{(t)})\right), \quad \pi_t(x_i^{(t)}) = K_t x_i^{(t)} + b_t. \tag{3.44}$$

Via a recursion we can obtain the samples after τ timesteps, $\{x_i^{(t+\tau)}(K_t, b_t)\}_i$, where we make the dependence on K_t and b_t explicit. We assume that the same controller (K_t, b_t) is used at each time forward. By applying Algorithm 3.1, we can obtain the distance estimate $D_{H,n_{\text{values}}}\left(\{x_i^{(t+\tau)}(K_t, b_t)\}_i, \mu_Y\right)$, which is a differentiable function of the controller (K_t, b_t) (using the sigmoid trick). Therefore, to select the controller for time t, we start with an initial controller and apply gradient descent with step-size ρ to improve our controller design:

$$\pi_t \leftarrow \pi_t - \rho \nabla_{K_t, b_t} D_{H,n_{\text{values}}}\left(\{x_i^{(t+\tau)}(K_t, b_t)\}_i, \mu_Y\right). \tag{3.45}$$

Although the method to compute the controllers is presented in Equation (3.45) using simple gradient descent, we can apply more advanced descent algorithms, such as RMSProp [62], Adagrad [63], and Adam [64]. We can use multiple restarts and other empirical techniques to improve the robustness and convergence of the optimization process. The ability to apply any of these methods however relies on the fact that our distance metric in the space of distributions is an easily computable and differentiable function of the controller. Here, we only consider simple gradient descent with a fixed step-size ρ and a fixed number of iterations. We leave the exploration of more advanced optimization methods to future work.

We run variants of our algorithm (3.45), where we vary H and n_{values}. However, we judge the performance of all variants by looking at the estimate $D_{300,100}$ for the distance between the state distribution and the target over time. We use $N = 3000$. For all $D_{H,n_{\text{values}}}(\cdot, \mu_Y)$, for each q_k, we use equally spaced values between the 0.5th and 99.5th percentiles of $q_k^\top Y$, where $Y \sim \mu_Y$, for the b_js. We set $\tau = 25$ and only update the controller every 20 steps.

We evaluate our algorithm on the unicycle dynamics model, which is nonlinear and given by the equation

$$\begin{bmatrix} \chi_{t+1} \\ \psi_{t+1} \\ \theta_{t+1} \end{bmatrix} = \underbrace{\begin{bmatrix} \chi_t + u_{1,t}\Delta t \cos(\theta_t) \\ \psi_t + u_{1,t}\Delta t \sin(\theta_t) \\ \theta_t + u_{2,t}\Delta t \end{bmatrix}}_{f}, \tag{3.46}$$

where $u_{1,t}$ is the linear velocity input and $u_{2,t}$ is the angular velocity input. We will only look at the distribution of the Cartesian position (χ, ψ). We use $\Delta t = 0.1$. We will assume zero initial heading and

$$\mu_{\text{start}} = \mathcal{N}\left(\begin{bmatrix} -2 \\ -2 \end{bmatrix}, I\right), \; \mu_Y = \mathcal{N}\left(\begin{bmatrix} 3 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 & 1.5 \\ 1.5 & 2 \end{bmatrix}\right) \tag{3.47}$$

for the distributions of the position. We do not constrain the distribution of the heading at the final time and the initial heading is 0.

Our results are summarized in Figure 3.9. We observe that our algorithm (3.45) is sufficient to steer the state distribution of the unicycle model. We cannot exactly reach the target distribution, because the unicycle dynamics are non-linear, the controller is affine, and gradient descent is not guaranteed to converge to the global optimum. The oscillations for some variants in Figure 3.8 are due to the unicycle dynamics and the approximation error of the distance estimate for these variants in (3.45).

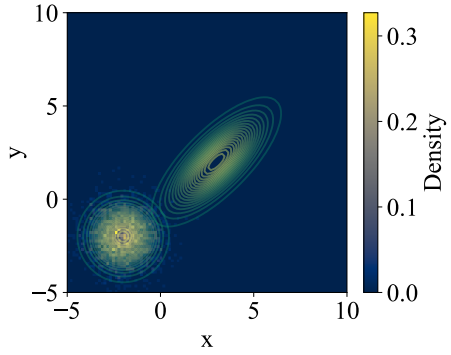


Figure 3.6: The sample density for the distribution of the initial position. The contours of the initial position distribution, μ_{start}, and the target distribution, μ_Y, are also included.

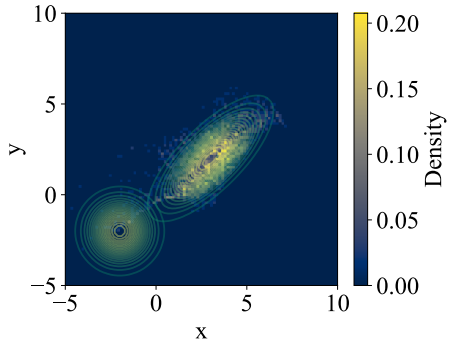


Figure 3.7: The sample density for the distribution of the final position for the algorithm variant of (3.45) with $H = 200$ and $n_{\text{values}} = 400$. Our proposed method steers the position distribution onto the target distribution.

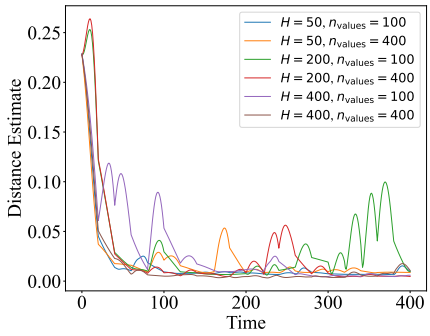


Figure 3.8: Running distance estimate $D_{300,100}(\cdot, \mu_Y)$ between the sample distribution of positions at time t and the target position distribution, for different variants of algorithm (3.45).

Figure 3.9: Results for the distribution steering application.

3.8 Ergodic control

The goal of ergodic control is to control the empirical distribution of a dynamical system's state as given by its trajectory over time [34], [35]. This is different than the goal of distribution steering, where we control the distribution of the dynamical system's state at any given time. We formulate the ergodic control problem as follows. We consider the dynamics (3.43) and select the control sequence $u =: (u_0, \dots, u_T)$, such that the distance

$$D_{H,n_{\text{values}}}\left(\{x_t\}_{t=0}^{T+1}, \mu_Y\right) \tag{3.48}$$

is minimized, where $x_{t+1} = f(x_t, u_t)$. Again, we can solve this problem using gradient descent

$$u \leftarrow u - \rho \nabla_u D_{H,n_{\text{values}}}\left(\{x_t\}_{t=0}^{T+1}, \mu_Y\right) \tag{3.49}$$

or a more advanced descent algorithm.

We assume the dynamics are unicycle, the initial state is $x_0 = (0, 0, 0)$, T=5,000, and the target position distribution μ_Y is a GMM of two components. We do not constrain the distribution of the heading trajectory. We run variants of (3.49) with different H and n_{values}. However, we evaluate the performance of all variants using $D_{400,400}$. We use AdamW [65], instead of the simple gradient descent update in (3.49). For all $D_{H,n_{\text{values}}}(\cdot, \mu_Y)$, for each q_k, we use equally spaced values between the 0.5th and 99.5th percentiles of $q_k^\top Y$, where $Y \sim \mu_Y$, for the b_js.

Our results are summarized in Figure 3.12. If we look at the position trajectory $x_0, \dots, x_{T+1}$ as a set of samples, we notice that the empirical distribution is close to the target distribution μ_Y. This is also verified by the loss at every gradient descent step that decreases until it reaches a point close to 0. The loss at every optimization step is the distance of the position trajectory at that step to the target distribution.

As a baseline, we also compare our method to simply maximizing the log-likelihood of the trajectory under the target distribution μ_Y. We use gradient descent here as well. The obtained trajectory is shown in Figure 3.13. We observe that the empirical distribution of the position trajectory is not close to the target distribution as it misses the second mode of the target distribution. This is because the log-likelihood

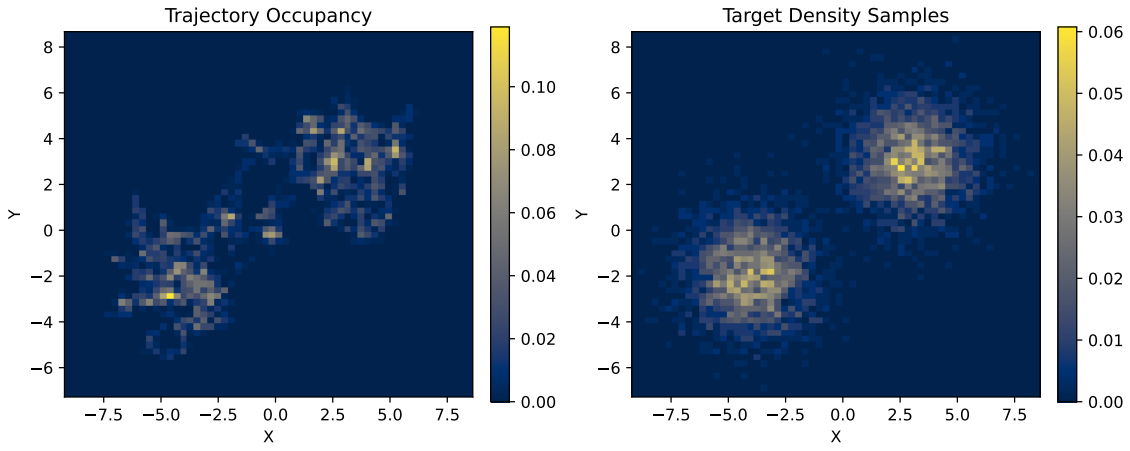


Figure 3.10: The state trajectory for the position, obtained by our algorithm (3.49) and represented as a sample density on the left, and the target GMM position density on the right. We observe that the empirical distribution of the position trajectory is close to the target distribution. On the left, the trajectory was obtained by the variant of algorithm (3.49) with $H = 50$ and $n_{\text{values}} = 400$. All other variants also get close to the target distribution, as indicated by Figure 3.11.

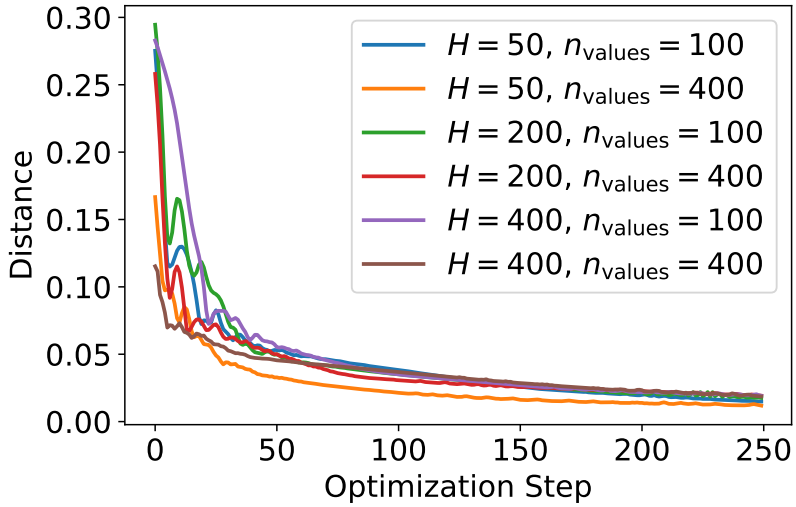


Figure 3.11: Distance estimate $D_{400,400}\left(\{x_t\}_{t=0}^{T+1}, \mu_Y\right)$ of the position trajectory to the target distribution as a function of optimization step for variants of algorithm (3.49).

Figure 3.12: Results for the ergodic control application.

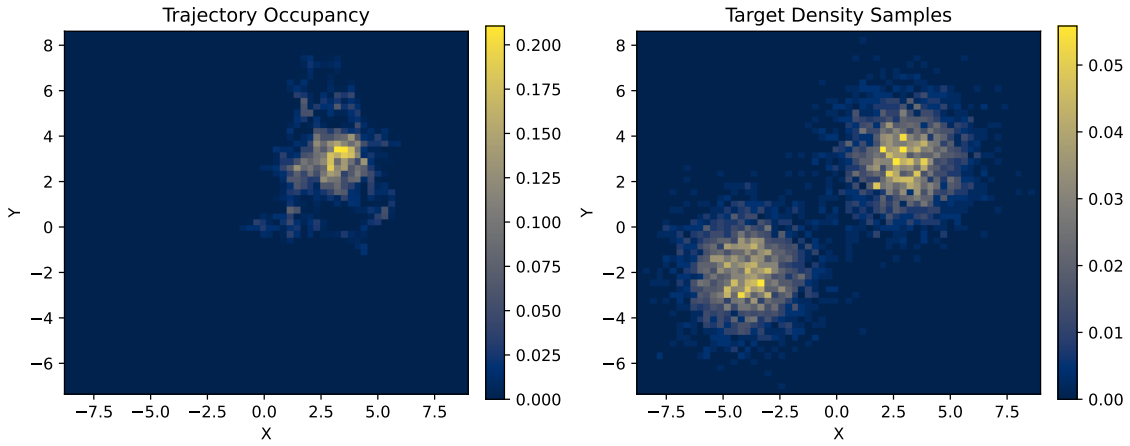


Figure 3.13: The state trajectory for the position, obtained by maximizing the log-likelihood under the target distribution μ_Y, and represented as a sample density on the left, and the target GMM position density on the right. We observe that the empirical distribution of the position trajectory is not close to the target distribution: the trajectory misses the second mode of the target distribution.

objective cannot capture the higher-order structure of the target distribution, while our proposed distance can capture this structure.

3.9 *Discussion*

We propose a family of distances between distributions that is interpretable and admits an easy-to-compute and differentiable approximation. We further present proof-of-concept applications of the proposed distance to control tasks.

Future work will focus on the finite-sample convergence behavior of Algorithm 3.1 as a function of the problem dimensionality and the sensitivity of Algorithm 3.1 to the selection of half-spaces. A principled way to select maximally discriminative half-spaces in Algorithm 3.1 is of particular interest. We will also look at the performance of Algorithm 3.1 in more complex control tasks, such as multi-agent control and control with partial observability.

An important application of the proposed distance is in training and evaluating generative models. The distance can be used as a loss function for training generative

models, and as a metric for evaluating the quality of generated samples. We will look at the performance of the proposed distance in these applications and compare it to other commonly used loss functions, such as the Kullback-Leibler divergence. Connections with conformal prediction and conformal inference [66], [67], [68] are also potentially fruitful.

4 *Informational representation*

In Chapter 3 we introduced a family of projection-based distances between probability distributions that is computationally tractable and allows for expressive control and planning.

In this chapter, we shift our focus back to a model of uncertainty based only on statistical moments. While control objectives that depend on the full distribution can capture richer features of uncertainty, moment-based descriptions are still preferable in practice because they are simple to estimate, computationally efficient in high dimensions, and easily compatible with computational tools for optimization and control.

In Chapter 2 we showed how to use the available samples from a distribution to estimate its covariance. We then showed how one can use this covariance model to solve the downstream task of portfolio optimization. If no other information except samples form the distribution is available, then computing the statistical moments is the best we can do. However, in many cases, we have access to additional variables other than the variable of interest. For example, in the case of financial returns, we have access to various economic indicators that are correlated with the returns. In the case of demand forecasting, we have access to past demands, which are correlated with future demands. We can then learn a joint probabilistic model for the random variable of interest and the additional information, and use it to *condition* on the additional information to get a better estimate of the moments of the random variable of interest. We call this idea the *informational* representation of uncertainty, as we use the information from the additional variables to get a better moment estimate of the random variable of interest. In other words, we

project the model of uncertainty on a family of models that is consistent with the observed values of the additional information. Not all statistical models allow one to easily condition on additional information. Selecting a model that fits the data and allows for easy conditioning is important for tractable estimation and control. In this chapter, we show how to use the informational representation of uncertainty to solve a sequential resource allocation problem under stochastic demands.

Furthermore, although conditioning is a well-known tool in probability, in this chapter we demonstrate how it can be used in the context of sequential decision-making. We show how to use the informational representation of uncertainty to solve the problem of optimally allocating a limited number of resources across time to maximize revenue under stochastic demands. This formulation is relevant in various areas of control, such as supply chain, ticket revenue maximization, healthcare operations, and energy allocation in power grids. We propose a bisection method to solve the optimization problem and extend our approach to a shrinking horizon algorithm for the sequential problem. The shrinking horizon algorithm computes future allocations after updating the distribution of future demands by conditioning on the observed values of demand. We illustrate the method on a simple synthetic example with jointly log-normal demands, showing that it achieves performance close to a bound obtained by solving the prescient problem.

4.1 Introduction

The problem of optimally allocating a finite quantity of resources over a discrete time horizon to satisfy stochastic demands is a fundamental challenge in control theory and operations research. This formulation can be viewed as a multi-period generalization of the classic newsvendor problem [69], extending the single-period inventory decision to a sequence of interdependent choices. The practical applications of this model are extensive and span areas such as supply chain management [70], dynamic pricing with fixed capacity [71], operational planning in healthcare [72], and the dispatch of energy resources in power grids [73]. In each case, the objective is to maximize the total revenue, subject to a constraint on the total available resources.

In this chapter, we first analyze the static multi-period resource allocation problem. In this formulation, the allocation decisions for the entire horizon are determined at the beginning of time. This approach yields a convex optimization problem, for which we can find an efficient solution using a bisection method on the dual variable. However, the solution to the static problem is inherently limited; it relies solely on the marginal distribution of demands for each time period and fails to incorporate temporal correlations or the information revealed as demands are sequentially realized. It is an open-loop policy that does not adapt to observed outcomes.

To overcome these limitations, we propose a sequential, adaptive policy based on a shrinking horizon optimization framework, similar to model predictive control (MPC) [74]. Under this policy, at every time, an allocation decision is made only for the current time. At each subsequent time, the optimization problem is re-solved for the remaining (shrinking) time horizon. Crucially, the joint distribution of future demands is updated by conditioning on the history of observed demands. This closed-loop approach allows the allocation strategy to dynamically adapt to new information as it becomes available. While this sequential method is a heuristic and does not guarantee global optimality, it is designed to systematically improve upon the static solution by leveraging the observed data.

The primary contribution of this chapter is the formulation and empirical evaluation of this shrinking horizon policy. We demonstrate its superiority against the static, open-loop solution. The evaluation is conducted on synthetic data where demands are drawn from a jointly log-normal distribution, a model chosen to reflect the correlated and non-negative nature of demands often found in practice. The results show that the adaptive method achieves higher revenue by effectively exploiting the information revealed over time. Its performance is close to a bound obtained by solving the prescient problem.

The remainder of this chapter is organized as follows. In Section 4.2 we discuss prior work. In Section 4.3 we describe the static optimization problem to maximize expected revenue and in Section 4.4 we show how to solve it. In Section 4.5 we describe the sequential problem of maximizing revenue and in Section 4.6 we

demonstrate how to obtain a policy for it. Finally, in Section 4.7 we describe our simulation set-up and results. Finally, we conclude the chapter in Section 4.8.

4.2 Related work

The static multi-period resource allocation problem is an extension of the classic newsvendor model [75]. We consider a portfolio of decisions constrained by a total resource limit. This class of problems has been extensively studied in network revenue management [76], but not under log-normal stochastic models. While foundational, the classic newsvendor model's assumption of a known demand distribution has been challenged by modern, data-rich environments. Recent advancements have thus focused on the data-driven newsvendor problem, where decisions are made directly from demand samples without assuming a specific parametric distribution [77], [78]. This line of work bridges the gap between classic stochastic models and modern adaptive methods, though often still focusing on a single-period setting [79]. Even when extended to multi-period settings, these static, open-loop policies are inherently suboptimal as they fail to adapt to the sequential arrival of information [80].

Adaptive policies have been proposed for the multi-period resource allocation problem. We mention some works that we believe capture the general trends of prior research. Kim et al. propose a policy assuming the uncertainty is expressed as a discrete set of possible scenarios. Contrary to our formulation, the random quantities do not appear in the optimization objective [81] and a discrete set of scenarios is used. Wang et al. propose an approximate dynamic programming approach under a complicated, partially observable information pattern [82]. Dynamic programming has been used to solve the problem under the assumption of independent and identically distributed demands [83]. Scenario trees, whose complexity scales exponentially with the number of nodes, have been proposed to solve discretized versions of the problem [84]. Approaches based on mixtures of experts have also been suggested [85]. Our method is different from the aforementioned works in the following ways: we propose and solve an abstraction of the sequential allocation

problem that is widely applicable. Our problem formulation is highly descriptive. We suggest a statistical model for demands that works well in practice and show how to incorporate it in an adaptive allocation scheme with good performance. Our algorithm is fast and scales well with the problem horizon.

Our shrinking horizon approach is a form of MPC, a well-known technique for approximating dynamic programming solutions. While MPC has been applied to supply chain problems, many implementations rely on certainty-equivalent formulations that use point forecasts [86]. Our method differs by directly incorporating the stochastic nature of the problem at each step, updating the statistical moments of future demands based on past observations. This allows our policy to explicitly react to temporal demand correlations, a significant challenge that often complicates other adaptive approaches.

4.3 *The static problem*

We consider a sequence of random variables $d_1, \ldots, d_T \in \mathbb{R}$, where T is the discrete-time horizon of the problem and the random variables represent the unknown demands at every time. A demand d_t means that d_t units or items were demanded at time t. Because demands are nonnegative, it always holds that

$$d_1, \ldots, d_T \geq 0. \tag{4.1}$$

We do not assume that the demands $d_1, \ldots, d_T$ are independent or identically distributed. In many applications, demands at different times are correlated and their marginal distributions can be significantly different.

For now, we assume that we have knowledge of the joint distribution of $d_1, \ldots, d_T$ and that for any time $t = 1, \ldots, T$, we can compute the marginal cumulative distribution function (CDF) $F_t(x) =: \mathbb{P}(d_t \leq x)$ for any $x \in \mathbb{R}$. Equivalently, we assume that we can compute the quantile functions $F_t^{-1}(y)$ for $y \in [0, 1]$ and $t = 1, \ldots, T$. This is possible, for example, when we can sample $(d_1, \ldots, d_T)$ from their joint distribution or when the joint distribution is known in closed form. We discuss

appropriate models for the distribution of the demands in Section 4.7.1.

The static resource allocation problem is

$$\begin{aligned} \max_{a_1,\ldots,a_T} \quad & \mathbb{E}_{d_1,\ldots,d_T}\left[\sum_{t=1}^{T} p_t \min\left(d_t, a_t\right)\right] \\ \text{s.t.} \quad & \sum_{t=1}^{T} a_t = L \\ & a_1, \ldots, a_T \geq 0 \end{aligned} \quad , \tag{4.2}$$

where the variables $a_1, \ldots, a_T \in \mathbb{R}_+$ are the allocated units or items for each time and $p_t > 0$ is the price per unit at time t. We are maximizing the expected revenue by optimally allocating the resources across time. The first constraint implies that there is a total of $L > 0$ units to be allocated across time and the second constraint does not allow for a negative allocation on any given time.

Problem (4.2) is a convex program. To see this, we proceed as follows. The objective of (4.2) can be rewritten as

$$\sum_{t=1}^{T} p_t \mathbb{E}_{d_t}\left[\min\left(d_t, a_t\right)\right], \tag{4.3}$$

where only the marginal distributions of the random variables, $d_1, \ldots, d_T$ are used. For every realization $\tilde{d}$, the function $\min\left(\tilde{d}, a_t\right)$ is concave in a_t. The expectation can be viewed as a sum with nonnegative weights of terms $\min\left(\tilde{d}, a_t\right)$ for different values $\tilde{d}$. Therefore, $\mathbb{E}_{d_t}\left[\min\left(d_t, a_t\right)\right]$ is concave in a_t. Because $p_t > 0$, we finally get that the function in (4.3) is concave in $(a_1, \ldots, a_T)$.

4.4 *The static solution*

We propose a method to solve problem (4.2). First, we find an equivalent expression for the objective. Then, we leverage duality to solve the problem.

4.4.1 An equivalent form for the objective of (4.2)

Consider the term $\mathbb{E}_{d_t}[\min(d_t, a_t)]$. We can write

$$\begin{aligned}\mathbb{E}_{d_t}[\min(d_t, a_t)] &= \int_0^{\infty} \mathbb{P}(\min(d_t, a_t) \geq x)\, dx \\ &= \int_0^{a_t} \mathbb{P}(\min(d_t, a_t) \geq x)\, dx \\ &= \int_0^{a_t} \mathbb{P}(d_t \geq x)\, dx.\end{aligned} \tag{4.4}$$

The first equality is a well-known result in probability [56]. The second equality follows because $\min(d_t, a_t) \leq a_t$. Finally, the third equality follows because for any $0 \leq x \leq a_t$

$$\min(d_t, a_t) \geq x \Leftrightarrow d_t \geq x. \tag{4.5}$$

By defining the survival functions

$$S_t(x) =: \mathbb{P}(d_t > x) = 1 - \mathbb{P}(d_t \leq x), \tag{4.6}$$

we can write the objective of (4.2) as

$$\sum_{t=1}^{T} p_t \int_0^{a_t} S_t(x) dx, \tag{4.7}$$

where the optimization variables a_t appear on the top limit of the integrals.

4.4.2 Optimality conditions

The Lagrangian of problem (4.2) is given by

$$\begin{aligned}&L(a_1, \ldots, a_T, \lambda_1, \ldots, \lambda_T, \nu) = \\ &\sum_{t=1}^{T} p_t \int_0^{a_t} \mathbb{P}(d_t \geq x)\, dx + \sum_{t=1}^{T} \lambda_t a_t - \nu \left(\sum_{t=1}^{T} a_t - L \right),\end{aligned} \tag{4.8}$$

where ν is the dual variable for the resource limit constraint and, for every t, $\lambda_t \geq 0$ is the dual variable for the non-negativity constraint $a_t \geq 0$.

The Karush–Kuhn–Tucker (KKT) optimality conditions [87] dictate that the primal-dual optimal point, $(a_1^\star, \ldots, a_T^\star, \lambda_1^\star, \ldots, \lambda_T^\star, \nu^*)$, satisfies

$$p_t \frac{\partial}{\partial a_t} \left(\int_0^{a_t} \mathbb{P}\left(d_t \geq x\right) dx \right) |_{a_t^\star} + \lambda_t^\star = \nu^\star, \tag{4.9a}$$

$$\sum_{t=1}^{T} a_t^\star = L, \tag{4.9b}$$

$$\lambda_t^\star a_t^\star = 0, \tag{4.9c}$$

$$\lambda_t^\star \geq 0, \quad a_t^\star \geq 0. \tag{4.9d}$$

Using the Leibniz integral rule [88], the first condition (4.9a) can be written as

$$p_t \mathbb{P}\left(d_t \geq a_t^\star\right) + \lambda_t^\star = \nu^\star. \tag{4.10}$$

Eq. (4.10) offers an intuitive explanation for the solution $(a_1^\star, \ldots, a_T^\star)$. For times t that the price per item, p_t, is high, the allocation $a_t^\star$ is large. To see this, simply ignore $\lambda_t^\star$. However, the relationship between price p_t and allocation $a_t^\star$ is not linear. Rather, it depends on the distribution of d_t.

We can obtain an expression for $a_t^\star$ as a function of $\nu^\star$. We suppose that for any $\delta > 0$: $\mathbb{P}\left(d_t \leq \delta\right) > 0$.

- Suppose $a_t^\star > 0$. Then, by (4.9c), $\lambda_t^\star = 0$ and, by (4.10), we get that $\mathbb{P}\left(d_t \geq a_t^\star\right) = \nu^\star / p_t < 1$, by the assumption above.
- Suppose $a_t^\star = 0$. Then $\mathbb{P}\left(d_t \geq a_t^\star\right) = 1$ and, since $\lambda_t^\star \geq 0$, by (4.10), we get that $\nu^\star / p_t \geq \mathbb{P}\left(d_t \geq 0\right)$, i.e., $\nu^\star / p_t \geq 1$.

Therefore, we obtain that

$$a_t^\star = \begin{cases} F_t^{-1}(1 - \dfrac{\nu^\star}{p_t}), & \dfrac{\nu^\star}{p_t} < 1, \\ 0, & \text{otherwise} \end{cases}, \tag{4.11}$$

where we can obtain $\nu^\star$ through (4.9b). Obviously by (4.10), it holds that $\nu^\star \geq 0$.

Notice from (4.11) that $\sum_{t=1}^{T} a_t^\star$ is a decreasing function of $\nu^\star$. Furthermore, suppose that for every t, it holds that $\nu^\star / p_t \geq 1$. Then $a_t^\star = 0$ for all t, which is not possible because $L > 0$. Therefore, there exists a t such that $\nu^\star / p_t < 1$, which implies

$$\nu^\star \in [0, \max\{p_1, \ldots, p_T\}]\,. \tag{4.12}$$

We can solve the system (4.9b)-(4.11) using bisection on $\nu^\star$. We can use 0 and $\max\{p_1, \ldots, p_T\}$ as the initial lower and upper limits for $\nu^\star$, respectively. This procedure is outlined in Algorithm 4.1. For an accuracy of ϵ for $\nu^\star$, we require $\mathcal{O}\left(\log\left(\max\{p_1, \ldots, p_T\}/\epsilon\right)\right)$ iterations. The only information required about the distribution of the demands are the quantile functions $F_t^{-1}(\cdot)$. These are straightforward to compute for various distributions, such as Gaussian or Laplace. Even for Gaussian mixture models, this is easy using a root-finding algorithm [89]. We call the allocations produced by Algorithm 4.1 the *static* solution.

In some applications, the allocations need to be integer-valued. We do not impose such a constraint in problem (4.2). However, after obtaining the solution $(a_1^\star, \ldots, a_T^\star)$ using Algorithm 4.1, we can round the allocation at each time to the closest integer and slightly modify them to ensure they sum to L. We can also easily do a local optimization around the rounded values. We note that with these slight modifications the duality gap is still bounded by a small number and not by a factor growing with T because there is only one equality constraint [90], [91].

It is important to note that we can solve problem (4.2) using Algorithm 4.1 at the beginning of time and then allocate the optimal number of units $a_1^\star, \ldots, a_T^\star$ over time. However, this approach can be sub-optimal because it only uses the marginal distributions of $d_1, \ldots, d_T$ and ignores the correlation between the demands across time. Furthermore, the demands $d_1, \ldots, d_T$ are realized sequentially, which Algorithm 4.1 does not leverage to improve performance.

Algorithm 4.1 Solution of the static resource allocation problem (4.2) using bisection

Inputs: horizon T, prices per item $p_1, \ldots, p_T > 0$, resource allocation limit $L > 0$, CDFs $F_t(u) = \mathbb{P}(d_t \leq u)$ for all $t = 1, \ldots, T$
Parameters: accuracy ϵ for solution $\nu^\star$
$N_{\text{iterations}} \leftarrow \lceil \log(\max\{p_1, \ldots, p_T\}/\epsilon) \rceil$
$\nu_{\text{low}} \leftarrow 0$
$\nu_{\text{up}} \leftarrow \max\{p_1, \ldots, p_T\}$
for $k = 1, \ldots, N_{\text{iterations}}$
    Obtain the midpoint $\nu_{\text{mid}} \leftarrow \dfrac{\nu_{\text{low}} + \nu_{\text{up}}}{2}$
    Obtain $\hat{a}_1, \ldots, \hat{a}_T$ through (4.11)
    if $\sum_{t=1}^{T} \hat{a}_t \geq L$
        $\nu_{\text{low}} \leftarrow \nu_{\text{mid}}$
    else
        $\nu_{\text{up}} \leftarrow \nu_{\text{mid}}$
return optimal allocations $\hat{a}_1, \ldots, \hat{a}_T$, optimal dual variable ν_{mid}

4.5 *The sequential problem*

We consider the case where the allocations need not all be determined at the beginning of time. At any time t, we assume that we have observed the demands $d_1, \ldots, d_{t-1}$ and already made the allocations $a_1, \ldots, a_{t-1}$, which cannot be altered. Our goal at time t is to determine the allocation a_t using the available information, which includes the realized demands, $d_1, \ldots, d_{t-1}$, the allocations made so far, $a_1, \ldots, a_{t-1}$, and the joint distribution of the demands $d_1, \ldots, d_T$. We assume that, for any time $t = 1, \ldots, T$ we have access to the marginal CDFs and quantile functions for the demands, conditioned on the previously observed values, i.e., $1, \ldots, t-1$ for every t. This is possible if we have selected a statistical model for the demands that allows for easy conditioning, i.e., an informational representation. We discuss appropriate models for the distribution of the demands in Section 4.7.1.

4.6 The sequential policy

We propose a policy based on shrinking horizon optimization for the sequential resource allocation problem. Our policy takes into account the realized demands so far to update the distribution of future demands through conditioning. More precisely, at time 0, we solve problem (4.2) to obtain the solution $(a^\star_{1|0}, \ldots, a^\star_{T|0})$, where the subscript denotes that the solution does not depend on any observed values of demand, since none have been observed yet.

After assigning the allocation $a^\star_{1|0}$ for time 1, we observe the realized value for the demand d_1. To compute the future allocations, we now solve

$$\begin{aligned} \max_{a_2,\ldots,a_T} \quad & \mathbb{E}_{d_2,\ldots,d_T|d_1}\left[\sum_{t=2}^{T} p_t \min\left(d_t, a_t\right)\right] \\ \text{s.t.} \quad & \sum_{t=2}^{T} a_t = L - a^\star_{1|0} \\ & a_2, \ldots, a_T \geq 0 \end{aligned} \quad , \tag{4.13}$$

which has the solution $(a^\star_{2|1}, \ldots, a^\star_{T|1})$. In the objective of (4.13) we have conditioned on the observed value of the demand. We assign the allocation $a^\star_{2|1}$ at time 2 and then observe the realized demand d_2. To make our approach clear, we mention that at time 3, we will solve

$$\begin{aligned} \max_{a_3,\ldots,a_T} \quad & \mathbb{E}_{d_3,\ldots,d_T|d_1,d_2}\left[\sum_{t=3}^{T} p_t \min\left(d_t, a_t\right)\right] \\ \text{s.t.} \quad & \sum_{t=3}^{T} a_t = L - a^\star_{1|0} - a^\star_{2|1} \\ & a_3, \ldots, a_T \geq 0 \end{aligned} \quad . \tag{4.14}$$

Our procedure is outlined in Algorithm 4.2. At each time t, we can use the bisection method in Algorithm 4.1 to obtain the allocations. Algorithm 4.2 requires knowing the conditional quantile functions for the demands, conditioned on the previously observed values. As in Section 4.4, we can round the allocation at each

time to the closest integer, if this is required by the application. Our approach is a sophisticated heuristic and need not be optimal. We call the allocations produced by Algorithm 4.2, the *sequential* solution.

As a side note, in the series of problems (4.13)-(4.14), we use the conditional mean in the objective. We could have also used the maximum a posteriori (MAP) value. Our policy is a heuristic and which option works best depends on the application.

4.6.1 The case of independent demands

In the case of independent demands $d_1, \ldots, d_T$, conditioning on realized demands does not alter the distribution of future demands. For example, the marginal distribution of $d_2, \ldots, d_T$ is the same as the conditional distribution $d_2, \ldots, d_T \mid d_1$. This implies that

$$(a^\star_{2|0}, \ldots, a^\star_{T|0}) = (a^\star_{2|1}, \ldots, a^\star_{T|1}). \tag{4.15}$$

To see why (4.15) is true, let

$$\begin{aligned} R_{2:} =: \mathbb{E}_{d_2,\ldots,d_T}\left[\sum_{t=2}^{T} p_t \min\left(d_t, a^\star_{t|0}\right)\right] \\ = \mathbb{E}_{d_2,\ldots,d_T|d_1}\left[\sum_{t=2}^{T} p_t \min\left(d_t, a^\star_{t|0}\right)\right]. \end{aligned} \tag{4.16}$$

The equality follows from independence. We suppose, for the sake of contradiction, that there exists an elementwise non-negative $(a_2, \ldots, a_T)$ such that

$$\mathbb{E}_{d_2,\ldots,d_T}\left[\sum_{t=2}^{T} p_t \min\left(d_t, a_t\right)\right] > R_{2:}, \tag{4.17a}$$

$$\sum_{t=2}^{T} a_t = L - a^\star_{1|0}. \tag{4.17b}$$

Algorithm 4.2 Shrinking horizon procedure for the sequential resource allocation problem (A.1)

1: **Inputs:** horizon T, prices per item $p_1, \ldots, p_T > 0$, resource allocation limit $L > 0$
2: Solve problem (A.1) using Algorithm 4.1 to obtain $(a^\star_{1|0}, \ldots, a^\star_{T|0})$
3: Assign the allocation $a^\star_{1|0}$ at time 1 and observe the realized demand d_1
4: **for** $\tau = 2, \ldots, T$
5: Solve problem

$$\begin{aligned} \max_{a_\tau, \ldots, a_T} \quad & \mathbb{E}_{d_\tau, \ldots, d_T | d_1, \ldots, d_{\tau-1}} \left[\sum_{t=\tau}^{T} p_t \min\left(d_t, a_t\right) \right] \\ \text{s.t.} \quad & \sum_{t=\tau}^{T} a_t = L - \sum_{t=1}^{\tau-1} a^\star_{t|t-1} \\ & a_\tau, \ldots, a_T \geq 0 \end{aligned} \tag{4.20}$$

using Algorithm 4.1 to obtain $(a^\star_{\tau|\tau-1}, \ldots, a^\star_{T|\tau-1})$.
6: Assign the allocation $a^\star_{\tau|\tau-1}$ at time τ and observe the realized demand d_τ

Then, it holds that $(a^\star_{1|0}, a_2, \ldots, a_T)$ is a feasible point for problem (4.2) and obtains a larger objective value than $(a^\star_{1|0}, \ldots, a^\star_{T|0})$. However, this is a contradiction. Therefore, for any elementwise non-negative $(a_2, \ldots, a_T)$ that satisfies (4.17b) it holds that

$$\begin{aligned} & \mathbb{E}_{d_2, \ldots, d_T | d_1} \left[\sum_{t=2}^{T} p_t \min\left(d_t, a_t\right) \right] \leq \\ & \mathbb{E}_{d_2, \ldots, d_T | d_1} \left[\sum_{t=2}^{T} p_t \min\left(d_t, a^\star_{t|0}\right) \right] \end{aligned} \tag{4.18}$$

Therefore, (4.15) holds, assuming a unique optimal point. Using recursion, we can prove that for any $\tau = 1, \ldots, T-1$:

$$(a^\star_{\tau+1|0}, \ldots, a^\star_{T|0}) = (a^\star_{\tau+1|\tau}, \ldots, a^\star_{T|\tau}). \tag{4.19}$$

Thus, when the demands are independent, the *sequential* solution obtained by our shrinking horizon policy reduces to the *static* solution.

4.7 Simulations

We compare the performance of the *static* and *sequential* solutions on synthetic data. First, we describe the model used for the demands and the generation process for the data. Second, we describe an upper bound for the obtained revenue and a simple baseline. Third, we show the equivalence of the *static* and *sequential* solutions in the case of independent demands. Fourth, we demonstrate the superiority of the *sequential* solution in the case of non-independent demands. Finally, we examine performance under model mismatch, where the true distribution of the demands is different from the one used to compute the allocations.

4.7.1 Log-normal model for the demands

We model the demands $d_1, \ldots, d_T$ as jointly log-normal random variables. Equivalently, we model $\log d_1, \ldots, \log d_T$ as a jointly Gaussian random vector with mean μ and covariance Σ, i.e.,

$$(\log d_1, \ldots, \log d_T) \sim \mathcal{N}(\mu, \Sigma). \tag{4.21}$$

The log-normal model is particularly suitable to model demands, because demands are non-negative, empirically exhibit right-skewness, and are correlated across time. The correlations across time can be incorporated elegantly, because the log demands are modeled as jointly Gaussian.

As a demonstration, in Figure 4.1 we show a histogram of log demands sampled from $\mathcal{N}(0, 1)$ and a histogram of the corresponding demand samples. The log-normal distribution is right-skewed, which is a common property of demands in many applications.

We now describe how to compute probabilities of demand under the log-normal model. We note that because $\mathbb{P}(d_t \leq x) = \mathbb{P}(\log d_t \leq \log x)$,

$$F_t(x) = \Phi\left(\frac{\log x - \mu_t}{\sqrt{\Sigma_{t,t}}}\right), \tag{4.22}$$

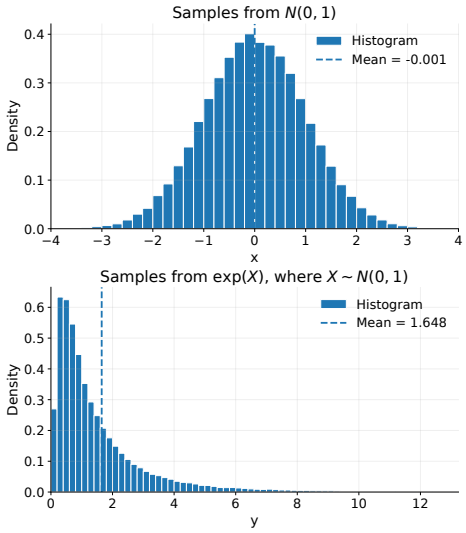


Figure 4.1: Histogram of log-demand samples drawn from $\mathcal{N}(0,1)$ and the corresponding demand samples obtained by exponentiation. The resulting log-normal demand distribution is right-skewed, a common characteristic of demands in many applications.

where μ_i is the i-th entry of vector μ, $\Sigma_{i,j}$ is the (i,j)-th element of matrix Σ, and $\Phi(\cdot)$ is the CDF of the standard normal distribution. The conditional distribution of d_t given $d_{t-1}, \ldots, d_1$ can be written as

$$\begin{aligned} &\mathbb{P}\left(d_t \leq x \mid d_{t-1} = x_{t-1}, \ldots, d_1 = x_1\right) = \\ &\mathbb{P}\left(\log d_t \leq \log x \mid \log d_{t-1} = \log x_{t-1}, \ldots, \log d_1 = \log x_1\right), \end{aligned} \tag{4.23}$$

because log is a monotone transformation. Assuming the conditional mean of $\log d_t$ is $\mu_{t|t-1}$ and its conditional standard deviation is $\sigma_{t|t-1}$, we get that

$$\begin{aligned} &\mathbb{P}\left(d_t \leq x \mid d_{t-1} = x_{t-1}, \ldots, d_1 = x_1\right) = \\ &\qquad \Phi\left(\frac{\log x - \mu_{t|t-1}}{\sigma_{t|t-1}}\right). \end{aligned} \tag{4.24}$$

We can compute $\mu_{t|t-1}$ and $\sigma_{t|t-1}$ explicitly using conditioning on jointly Gaussian distributions [57, Section 3.9].

Although modeling the log demands as jointly Gaussian is adequate for many

applications, we can model them using a Gaussian mixture model (GMM), which is a universal approximator [92]. In the case of the GMM, we can again efficiently evaluate the conditional quantile functions. Although we do not explore GMMs in our simulations, we include a discussion below.

Gaussian mixture models for log demands

For more expressivity, we can model the log demands as a Gaussian mixture model (GMM). A GMM is a weighted sum of K Gaussian distributions, where the weights π_i are non-negative and sum to 1. In this case

$$(\log d_1, \ldots, \log d_T) \sim \sum_{i=1}^{K} \pi_i \mathcal{N}(\mu_i, \Sigma_i). \tag{4.25}$$

We show a histogram of log demands sampled from a GMM and the corresponding demand samples in Figure 4.2. The resulting demand distribution is similar to simply using a single Gaussian distribution for the log demands. However, the GMM allows for more expressivity and can capture more complex distributions of the log demands.

Under a GMM model

$$F_t(x) = \sum_{i=1}^{K} \pi_i \Phi\left(\frac{\log x - (\mu_i)_t}{\sqrt{(\Sigma_i)_{t,t}}}\right). \tag{4.26}$$

This is a simple extension of the single Gaussian case.

Conditioning is also similar to the single Gaussian case. Let

$$y = \begin{bmatrix} y_{\text{obs}} \\ y_{\text{fut}} \end{bmatrix} = \begin{bmatrix} \log d_1 \\ \vdots \\ \log d_{t-1} \\ \log d_t \\ \vdots \\ \log d_T \end{bmatrix} \sim \sum_{i=1}^{K} \pi_i \, \mathcal{N}(\mu_i, \Sigma_i),$$

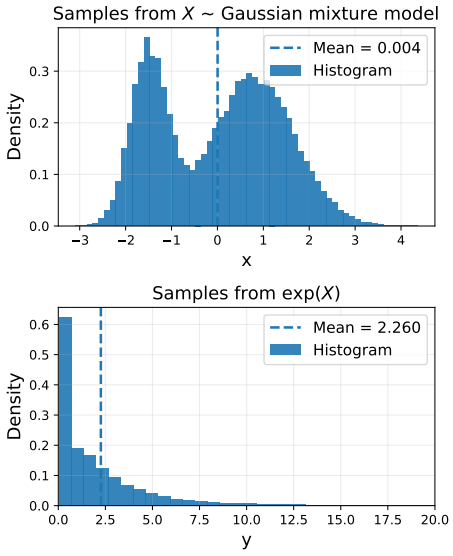


Figure 4.2: Samples from a Gaussian mixture model for log demand and the corresponding demand samples obtained through the transformation $d = \exp(\log d)$.

where

$$y_{\text{obs}} = \log d_{1:t-1}, \qquad y_{\text{fut}} = \log d_{t:T}.$$

For each mixture component i, partition the mean and covariance as

$$\mu_i = \begin{bmatrix} \mu_{i,\text{obs}} \\ \mu_{i,\text{fut}} \end{bmatrix}, \qquad \Sigma_i = \begin{bmatrix} \Sigma_{i,\text{oo}} & \Sigma_{i,\text{of}} \\ \Sigma_{i,\text{fo}} & \Sigma_{i,\text{ff}} \end{bmatrix}.$$

Given the observed realization

$$y_{\text{obs}} = (\log d_1, \ldots, \log d_{t-1}),$$

the conditional distribution remains a Gaussian mixture:

$$y_{\text{fut}} \mid y_{\text{obs}} \sim \sum_{i=1}^{K} \widetilde{\pi}_i(y_{\text{obs}})\, \mathcal{N}\Big(\mu_{i,\text{fut}|\text{obs}}, \Sigma_{i,\text{fut}|\text{obs}}\Big),$$

where

$$\mu_{i,\text{fut}|\text{obs}} = \mu_{i,\text{fut}} + \Sigma_{i,\text{fo}} \Sigma_{i,\text{oo}}^{-1} \left(y_{\text{obs}} - \mu_{i,\text{obs}}\right)$$

and

$$\Sigma_{i,\text{fut}|\text{obs}} = \Sigma_{i,\text{ff}} - \Sigma_{i,\text{fo}}\Sigma_{i,\text{oo}}^{-1}\Sigma_{i,\text{of}}.$$

These are the conditional mean and covariance of the multivariate Gaussian distribution corresponding to the i-th mixture component. They are computed the same way as $\mu_{t|t-1}$ and $\sigma_{t|t-1}$ in the single Gaussian case above.

The posterior mixture weights are

$$\widetilde{\pi}_i(y_{\text{obs}}) = \frac{\pi_i \mathcal{N}(y_{\text{obs}}; \mu_{i,\text{obs}}, \Sigma_{i,\text{oo}})}{\sum_{j=1}^{K} \pi_j \mathcal{N}(y_{\text{obs}}; \mu_{j,\text{obs}}, \Sigma_{j,\text{oo}})}.$$

For the next log demand $y_t = \log d_t$,

$$y_t \mid y_{1:t-1} \sim \sum_{i=1}^{K} \widetilde{\pi}_i \mathcal{N}\left(\mu_{i,t|t-1}, \sigma_{i,t|t-1}^2\right),$$

where

$$\mu_{i,t|t-1} = (\mu_{i,\text{fut}|\text{obs}})_1, \qquad \sigma_{i,t|t-1}^2 = (\Sigma_{i,\text{fut}|\text{obs}})_{1,1}.$$

Since $d_t = \exp(y_t)$, the conditional cumulative distribution function is

$$\mathbb{P}(d_t \leq x \mid d_1 = x_1, \ldots, d_{t-1} = x_{t-1}) = \sum_{i=1}^{K} \widetilde{\pi}_i \, \Phi\left(\frac{\log x - \mu_{i,t|t-1}}{\sigma_{i,t|t-1}}\right), \qquad x > 0.$$

4.7.2 *The synthetic data*

We describe the process of generating the synthetic data. We select prices p_t between 10 and 100. We select the means of the demands d_t between 20 and 100, while the standard deviations of the demands are between 10 and 30. The first moments of the demands are mapped appropriately to the means and standard deviations of the log demands using standard formulas [93].

The correlations between the log demands are selected differently in our experiments depending on whether the demands are independent or not. For the case of

independent demands we set the correlations to be 0 for log demands of different times. For the case of non-independent demands, we set the correlations between $\log d_t$ and $\log d_\tau$ to be (roughly) between -0.7 and 0.7 for $\tau \neq t$. To do so, we first create a symmetric matrix with diagonal entries equal to 1 and off-diagonal entries in $\{-0.7, 0.7\}$. We then project to the set of positive semi-definite matrices to obtain the correlation matrix.

We set the limit L to be between 30% and 60% of the sum of mean demands. The synthetic data is generated once and remains fixed across the simulation trials for each set of experiments below.

4.7.3 *A revenue upper-bound*

Suppose we know the realized sequence of demands, $\tilde{d}_1, \ldots, \tilde{d}_T$. Then, we can calculate the optimal allocations for this realization by solving the convex problem

$$\begin{aligned} \max_{a_1,\ldots,a_T} \quad & \sum_{t=1}^{T} p_t \min\left(\tilde{d}_t, a_t\right) \\ \text{s.t.} \quad & \sum_{t=1}^{T} a_t = L \\ & a_1, \ldots, a_T \geq 0 \end{aligned} \quad . \tag{4.27}$$

We call the solution of (4.27), the *oracle* allocation. This allocation cannot be computed in a causal manner and is only known after observing the realized demands. Nevertheless, it serves as a useful upper-bound of performance.

4.7.4 *The roll-forward baseline*

We describe a simple causal baseline. At time 1, we set the allocation to be L/T. At each subsequent time t, we set the allocation to be the last observed demand, i.e., $a_t = d_{t-1}$, if the remaining allocation is larger than d_{t-1}. Otherwise, we simply set a_t to be the remaining allocation. We call this baseline, the *roll-forward* allocation, because it just rolls the last observed demand forward as the next allocation. The

roll-forward method is a simple heuristic and serves as one of the simplest estimators of future demand. As a pracical use of the roll-forward method in finance, to make private company valuations, experts often use this method to estimate the current value of the company based on the last observed valuation and any cashflows in the meantime.

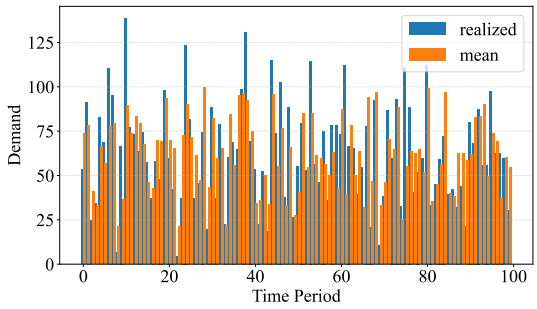
4.7.5 Independent demands

We set the correlation between $\log d_t$ and $\log d_\tau$ to be 0 for any $\tau \neq t$. This implies that the log demands are independent. As functions of the log demands, the demands are also independent in this case. By Section 4.6.1, we know that the *static* and *sequential* solutions must match. This is verified in our simulation, as shown in Figure 4.2(c), for a particular realization of demands, which is shown in Figure 4.2(a). In this case, the two approaches obtain the same revenue, which is smaller than the revenue obtained by the *oracle* allocation, as shown in Figure 4.2(d). However, the *oracle* allocation cannot be computed in a causal manner. The *roll-forward* method, which is causal, performs the worst, because it does not adequately capture the inter-dependencies between the demands. It also allocates the total L in far fewer time periods, which explains the constant cumulative revenue in later time periods. In Figure 4.2(c), we do not plot the *roll-forward* allocation because this can be easily inferred by the realized demands in Figure 4.2(a).

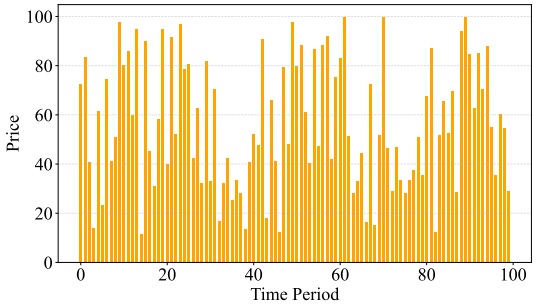
4.7.6 Non-independent demands

We set the correlations between $\log d_t$ and $\log d_\tau$ to be (roughly) between -0.7 and 0.7 for $\tau \neq t$. We first look at the performance of the methods for a single trial, i.e., a single realized sequence of demands, and then provide aggregate results over 100 trials for different time horizons T.

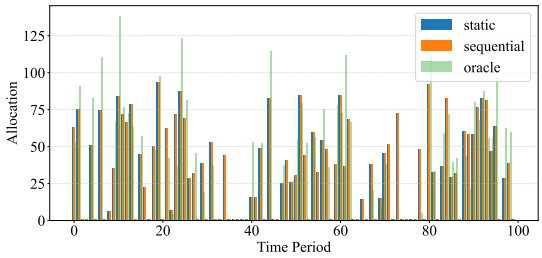
Our results for a single trial and $T = 100$ are included in Figure 4.4. We observe that the *sequential* allocation is closer to the oracle allocation than the *static* one. As a result, the cumulative revenue of the *sequential* allocation is higher than the cumulative revenue of the *static* one. We note that we expect the gap between the



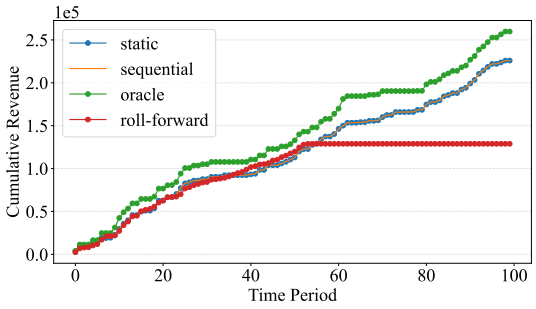
(a): Realized and expected demands across time.



(b): Prices (per unit) across time.



(c): The *oracle*, *sequential*, and *static* allocations across time. The *static* and *sequential* allocations are the same for independent demands.



(d): Cumulative revenue over time for the *oracle*, *sequential*, *static*, and *roll-forward* allocations. The cumulative revenue is the same for the *static* and *sequential* allocations.

Figure 4.3: Results for the case of independent demands.

revenue of the *sequential* and *static* allocations to shrink as the correlations between the demands get closer to 0, i.e., when the demands are close to being independent.

We include the aggregate results over 100 trials for different values of T in Table 4.1. Across values of T the *sequential* allocation outperforms the *static* one with respect to revenue. The *roll-forward* allocation does not have a good performance. We show the mean and the 1-standard deviation interval for the cumulative revenue of the different allocation schemes in Figure 4.5 for the different values of T.

Method	$T = 20$	$T = 50$	$T = 100$	$T = 200$
Roll-Forward	18,473 (1,429)	80,169 (5,154)	137,179 (6,221)	217,834 (6,804)
Static	36,644 (1,197)	118,164 (3,892)	227,780 (5,890)	386,890 (5,615)
Sequential	**39,426** (690)	**127,637** (4,199)	**249,393** (5,183)	**415,716** (4,713)
Oracle	41,185 (734)	132,933 (4,566)	256,823 (5,306)	425,021 (5,011)

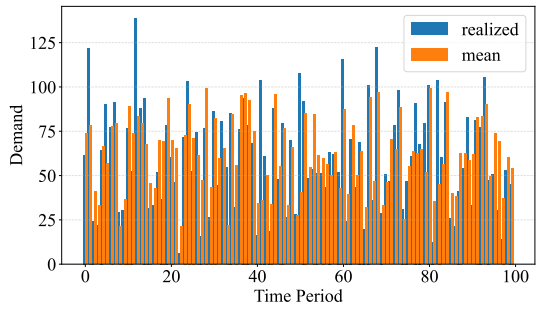
Table 4.1: Results across time horizons T for non-independent demands. Each cell shows two numbers (i.e., mean of achieved revenue on the first line and standard deviation of the achieved revenue in parentheses on the second line over 100 trials). In bold, we emphasize the causal method that is best in terms of revenue, i.e., the *sequential* allocation. We also observe that the *sequential* allocation tends to have the smallest revenue variance.

4.7.7 Model misspecification

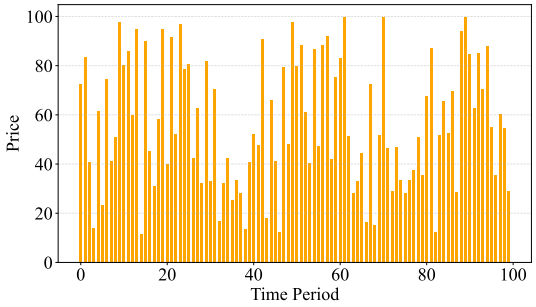
The experiments above assume that the probabilistic model used by the allocation methods coincides with the distribution used to generate the realized demands. In practice, the true demand distribution is unknown and must be estimated from data. The conditional distributions used by the *sequential* method can therefore differ from the true conditional distributions. Even the marginal distributions used by the *static* method can differ from the true marginal distributions. We study the sensitivity of the allocation methods to such modeling errors by distinguishing between the *real* demand model and the model available to the algorithms. Demand realizations are sampled from the real model, whereas the allocations are computed using a perturbed model.

We separately perturb the means, standard deviations, and correlations of the demand model. In the first experiment, the mean demand at each time is perturbed according to

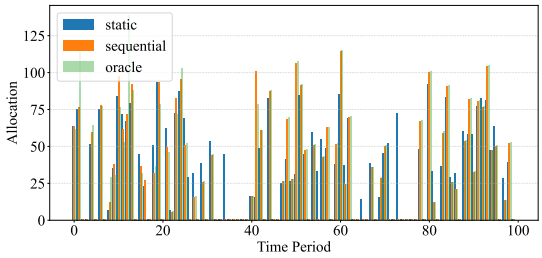
$$\mathbb{E}_{\text{model}}[d_t] = \mathbb{E}_{\text{real}}[d_t] + \epsilon_t, \qquad \epsilon_t \sim \mathcal{U}([-5, 5]). \tag{4.28}$$



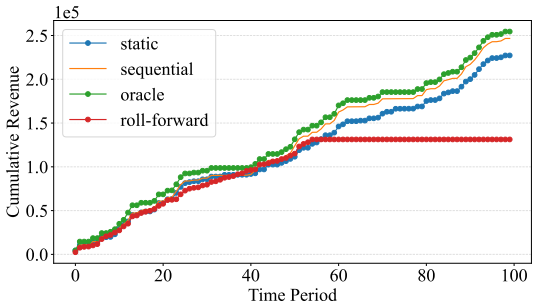
(a): Realized and expected demands across time.



(b): Prices (per unit) over time.



(c): The *oracle*, *sequential*, and *static* allocations across time. The *sequential* allocation is close to the *oracle* allocation.



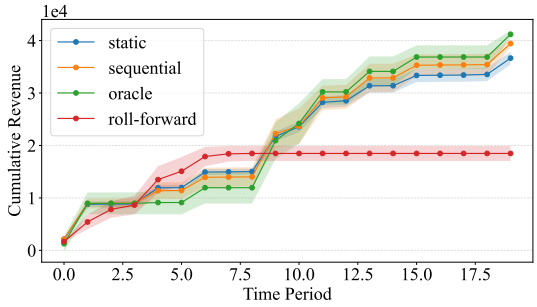
(d): Cumulative revenue over time for the *oracle*, *sequential*, *static*, and *roll-forward* allocations.

Figure 4.4: Results for the case of non-independent demands.

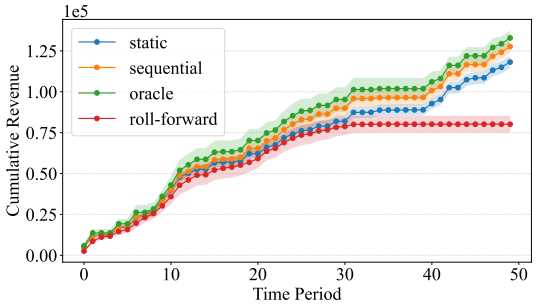
Real corresponds to the data generating process, while model corresponds to the model used by the allocation methods. This perturbation changes the demand levels anticipated by the allocation methods while leaving the distribution that generates the realized demands unchanged.

We next perturb the marginal standard deviations. We consider a moderate level of misspecification,

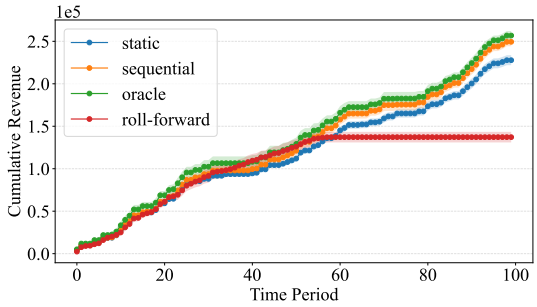
$$\text{std}_{\text{model}}(d_t) = \text{std}_{\text{real}}(d_t)\gamma_t, \qquad \gamma_t \sim \mathcal{U}([0.8, 1.2]), \tag{4.29}$$



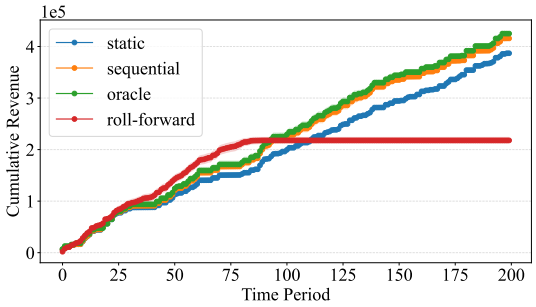
(a): $T = 20$.



(b): $T = 50$.



(c): $T = 100$.



(d): $T = 200$.

Figure 4.5: The mean and 1-standard deviation interval for the cumulative revenue over 100 trials for the various allocation methods and different horizons T.

and a stronger level,

$$\text{std}_{\text{model}}(d_t) = \text{std}_{\text{real}}(d_t)\gamma_t, \qquad \gamma_t \sim \mathcal{U}([0.5, 1.5]). \tag{4.30}$$

These cases test whether the methods remain effective when the model underestimates or overestimates the uncertainty associated with future demands.

Finally, we perturb the temporal dependence represented by the log-demand correlations. For each pair of times, the correlation used by the algorithms is given by

$$\text{corr}_{\text{model}}\left(\log d_\tau, \log d_t\right) = \text{corr}_{\text{real}}\left(\log d_\tau, \log d_t\right)\eta_{\tau t}, \qquad \eta_{\tau t} \sim \mathcal{U}([0.5, 1.5]). \tag{4.31}$$

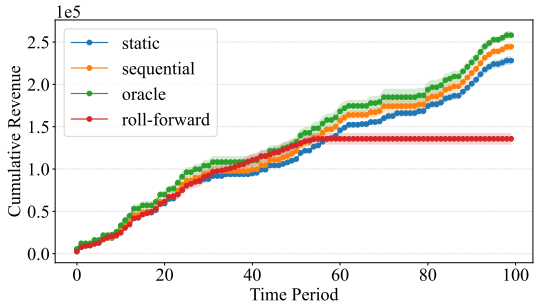
This experiment is especially relevant to the *sequential* method, because the conditional distribution of future demands depends directly on the modeled temporal correlations. An inaccurate correlation model can cause an observation to have either too much or too little influence on the allocations computed for the remaining horizon.

The results are shown in Figure 4.6. The *sequential* method continues to obtain higher revenue than the *static* and *roll-forward* methods across the different forms of misspecification. Its performance therefore does not rely on exact knowledge of the data-generating distribution. The degradation is more pronounced under stronger errors in the standard deviations and correlations, since these quantities influence both the conditional uncertainty and the extent to which past demands are informative about future demands. Nevertheless, the results indicate that the benefit obtained by conditioning and re-optimizing is retained under moderate discrepancies between the assumed and real demand models.

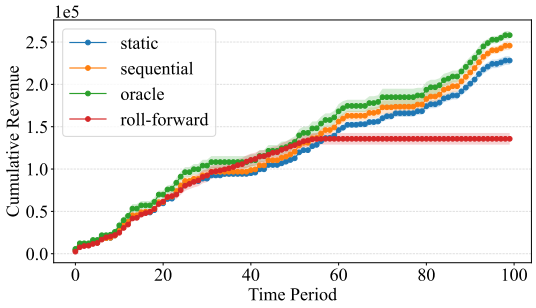
4.8 Discussion

We propose a general sequential decision-making algorithm to solve the multi-period resource allocation problem under stochastic demands. The approach leverages the sequential nature of the problem and the correlation between the demands across time to improve performance compared to a static solution, which does not use this information. This showcases the value of the *informational* representation of uncertainty. Our approach only requires knowledge of the conditional quantile functions. We show the superiority of our approach compared to the static solution using synthetic jointly log-normal data for the demands. The log-normal model is expressive and allows for easy calculation of the conditional quantile functions.

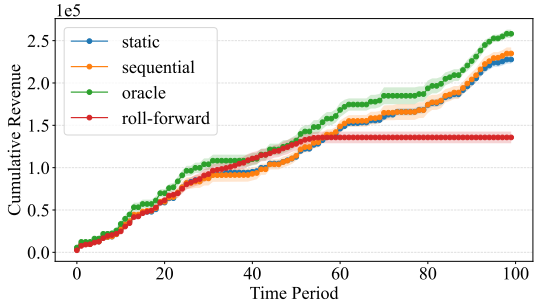
Future work will focus on incorporating Algorithm 4.2 in a closed loop, where the allocations are determined while the model for the distribution of the demands is learned. Furthermore, we plan to study in detail the importance of accurately modeling the distribution of the demands in applications. An interesting extension is accounting for uncertainty in the prices p_t in addition to the demands. Finally, we



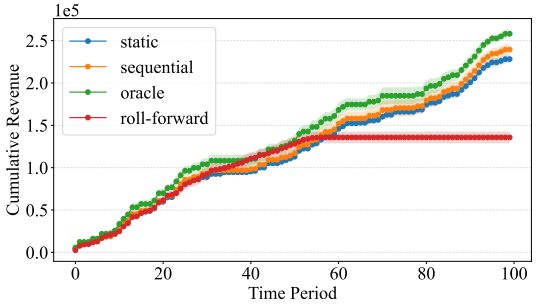
(a): Misspecification of the marginal demand means as in (4.28).



(b): Moderate misspecification of the marginal standard deviations as in (4.29).



(c): Stronger misspecification of the marginal standard deviations as in (4.30).



(d): Misspecification of the temporal log-demand correlations as in (4.31).

Figure 4.6: Cumulative revenue under different forms of model misspecification. Demand realizations are generated using the real model, whereas the allocations are computed using a model with perturbed means, standard deviations, or log-demand correlations.

will apply our approach to real data from a supply chain management application.

As a general remark, we hope that this chapter demonstrated the power of the informational representation of uncertainty (where conditioning is easy) combined with a model-predictive control (or shrinking horizon) approach to solve sequential decision-making problems under uncertainty. We believe that this approach can be applied to a wide range of problems beyond the resource allocation problem studied in this chapter.

Although Markov decision processes and partially observable Markov decision processes are powerful frameworks to model sequential decision-making problems under uncertainty, we find that a state estimation step (to update the statistical moments) combined with solving a model-predictive control problem (that can usually be formulated as a convex optimization problem) are sufficient to compute a control action at any given time and obtain good overall performance in many applications.

5 Summary and future work

Automated decision making is becoming increasingly prevalent in various domains, such as healthcare, finance, and autonomous systems. This is justified by the increase in computational capabilities, the abundance of data, and the efficiency, accuracy, and scalability that automated decision making can offer.

Although more data is available, leveraging it effectively remains a challenge. Data can be noisy, high-dimensional, and may only partly be relevant to the decision-making process. In this thesis, we discuss different represenations that capture the relevant information in the data and enhance the performance of automated decision making. These are in essence representations of the uncertainty that is inherent in the underlying problem and implied by the available data.

The proposed representations share a conceptual similarity with compression in information theory [94]: each replaces a high-dimensional or otherwise intractable object with a lower-complexity model that preserves the information most relevant. Unlike classical compression, however, the goal here is not faithful reconstruction of the original signal, but preservation of decision-relevant structure.

This chapter summarizes the modern decision making landscape, the contributions of this thesis, and potential avenues for future work.

5.1 Modern challenges in decision making

Modern decision-making problems arise in settings that are data-driven, high-dimensional, distribution-aware, and sequential. In such settings, the agent has

access to large amounts of data, but that data is often noisy and may contain spurious correlations. At the same time, both the state of the world and the decision variables may be high-dimensional, making it difficult to learn reliable models and to make good decisions. In many applications, the objective also depends on the full distribution of outcomes rather than only on low-order moments, which further complicates planning. Finally, decisions are made over time, and new information is revealed sequentially, so the agent must continually update both its model and its actions.

These challenges appear across many modern application domains. In the financial portfolio construction problem introduced in the introduction as a unifying example, the agent must estimate returns and covariances from historical data, make decisions over large asset universes, possibly optimize objectives that depend on tail events or target distributions, and repeatedly rebalance as market conditions evolve and new information arrives. Similar conditions arise more broadly in control, where methods must balance statistical fidelity, computational tractability, and responsiveness to new information.

5.2 *Contributions*

This thesis argues that choosing a suitable representation of the uncertainty inherent in the underlying problem is a key step in addressing the challenges of modern decision making: rather than attempting to model or optimize everything at full fidelity, one can seek lower-complexity representations that retain the information most relevant for the task at hand. By restricting attention to task-relevant structure, one can turn otherwise intractable problems into ones that are statistically meaningful and computationally manageable.

Three representations of uncertainty were introduced. Each was motivated by a different aspect of the modern decision-making landscape and was applied to a different problem domain, but all share the same underlying principle of preserving task-relevant information while reducing complexity. A chapter is dedicated to each representation.

- **Structural representation**. This representation was motivated by the need to learn tractable covariance models from data. In particular, we consider the family of low-rank-plus-diagonal covariance matrices, which are widely used in finance and control. We propose a method that takes a provided low-rank-plus-diagonal model and extends it with additional statistically learned factors while preserving tractability. We study this method in Chapter 2. It allows us to uncover hidden factors driving the behavior of the variables of interest.

- **Geometric representation**. This representation was motivated by the need to optimize distribution-aware control objectives. We propose a simple, computationally tractable method that compares distributions through lower-dimensional slices, enabling one to solve control problems with expressive distribution-aware control objectives. We cover this method in Chapter 3.

- **Informational representation**. This representation was motivated by the need to incorporate newly revealed information into sequential decision making. A decision making system should be able to handle regime shifts and adapt quickly to changing conditions. An informational representation of uncertainty is one that allows for easy updates through conditioning on observed data. We can then use the updated models to re-solve the control problem, allowing for a principled way to incorporate new information into sequential decision making. We cover this representation in Chapter 4. We show how to incorporate it in a closed-loop control scheme to solve sequential decision-making problems under uncertainty. Although conditioning is a well-known technique to incorporate information, we combine it with a shrinking horizon optimization approach for resource allocation under stochastic demands, which is a novel application domain.

5.3 *Future work*

A central direction for future work is to unify these three representations into a single framework for sequential decision making under uncertainty. In such a framework, the structural representation would focus on a tractable covariance model, the geometric representation would help express objectives over distributions, and the informational representation would allow for updating both the model and the decision as new observations arrive. Developing such an integrated framework would move beyond the separate case studies in this thesis toward a general methodology for modern decision making.

For the method extending the covariance model in Chapter 2, future work will focus on statistical, modeling, and computational questions. On the statistical side, it is important to better understand identifiability, finite-sample behavior, and the sensitivity of the learned extension to misspecification in the provided base model. On the modeling side, a key challenge is to select the number of additional factors in a principled way and to combine model extension with factor selection in the base model. On the computational side, promising directions include accelerating the EM procedure (see Section A.5 for a preliminary treatment), studying direct nonlinear optimization methods for low-rank-plus-diagonal model fitting, and developing online variants that update the model efficiently as new data arrive.

For the distance between distributions proposed in Chapter 3, future work will focus on the theoretical and algorithmic properties of the proposed family of distances between distributions. In particular, it is important to characterize finite-sample convergence as a function of dimension, the number of projection directions, and the discretization used within each slice. Another central question is how to select projection directions or half-spaces in a principled manner, ideally choosing those that are maximally informative for the task at hand. Beyond the proof-of-concept control examples studied here, the proposed distance should also be investigated in more complex settings such as partial observability, multi-agent control, and the training and evaluation of generative models.

For the shrinking horizon optimization approach in Chapter 4, the next step is to

move to closed-loop settings in which the distribution of uncertainty is not assumed known in advance, but is instead learned and updated online while actions are taken. Future work should study robustness to misspecification of the distribution of uncertainty and develop variants that explicitly account for uncertainty in the learned model. Applying the method to real data, for example in supply-chain management, is another important direction.

More broadly, this thesis suggests that simplification is not merely a computational convenience, but a useful design principle for modern decision making under uncertainty. The future challenge is to make this principle systematic: to understand when a representation of uncertainty is the simplest one that still preserves the information relevant for control, how can one select or design a representation based on the available data, and how can different representations be combined in larger pipelines, i.e., a learning-and-control loop.

Appendices

A Fitting a low-rank-plus-diagonal covariance model

We first (in Section A.1 to Section A.4) discuss properties of fitting a low-rank-plus-diagonal covariance model using a maximum likelihood approach. Then, in Section A.5, we discuss a variant of the EM algorithm that can speed up convergence by selecting the density used in the E-step in a principled way. Finally, in Section A.6, we present a way to perform the maximum likelihood fitting using a convex optimization package, although the problem is not convex in the variables.

A.1 Problem statement

We investigate the problem of fitting a low-rank-plus-diagonal model to a given covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$. Specifically, we will fit a matrix $F \in \mathbb{R}^{n\times p}$ and a diagonal matrix $D \in \mathbb{R}^{n\times n}$, by solving the following problem:

$$\underset{F,D}{\text{minimize}}\, \mathrm{KL}\left(\mathcal{N}(0,\Sigma)\|\mathcal{N}(0,FF^\top + D)\right). \tag{A.1}$$

The idea is that at optimality, for $\Sigma^\star =: F^\star F^{\star,\top} + D^\star$, where $F^\star$ and $D^\star$ are the optimal values, it holds that $\Sigma \approx \Sigma^\star$. We can equivalently write

$$\mathrm{KL}\left(\mathcal{N}(0,\Sigma)\|\mathcal{N}(0,FF^\top + D)\right) = \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\left[\log\frac{p(x)}{q(x)}\right], \tag{A.2}$$

where $p(x)$ is the probability density function (PDF) of $\mathcal{N}(0, \Sigma)$ and $q(x)$ is the PDF of $\mathcal{N}(0, FF^\top + D)$. Therefore, the solution of (A.1) is the same as the solution of

$$\underset{F,D}{\text{minimize}} \; -\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)} \left[\log q(x)\right], \tag{A.3}$$

where a log-likelihood objective appears and the problem can be written as

$$\underset{F,D}{\text{minimize}} \; \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)} \left[x^\top \left(FF^\top + D\right)^{-1} x + \log\det\left(FF^\top + D\right)\right]. \tag{A.4}$$

Using the fact that $\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)} \left[x^T A x\right] = \text{trace}\,(A\Sigma)$, we can write (A.4) as

$$\underset{F,D}{\text{minimize}} \; \text{trace}\left(\left(FF^\top + D\right)^{-1}\Sigma\right) - \log\det\left(FF^\top + D\right)^{-1}. \tag{A.5}$$

A.2 Fitting both F and D

We prove that for the solution, $F^\star$ and $D^\star$, of (A.5) (or equivalently (A.1)), under the constraint of diagonal D, it holds that

$$\text{diag}\left(F^\star F^{\star,\top} + D^\star\right) = \text{diag}\,\Sigma. \tag{A.6}$$

In other words, we prove that the implied asset volatilities of the fitted model match the volatilities of the original matrix.

First, by the Woodburry identity we get

$$\left(FF^\top + D\right)^{-1} = D^{-1} - D^{-1}F\left(I + F^\top D^{-1}F\right)^{-1} F^\top D^{-1} \tag{A.7}$$

and then we use the change of variables

$$\begin{aligned} \tilde{D} &= D^{-1} \\ \tilde{F} &= D^{-1}F\left(I + F^\top D^{-1}F\right)^{-1/2} \end{aligned} \quad \Longleftrightarrow \quad \begin{aligned} D &= \tilde{D}^{-1} \\ F &= \tilde{D}^{-1}\tilde{F}\left(I - \tilde{F}^\top \tilde{D}^{-1}\tilde{F}\right)^{-1/2} \end{aligned}$$

in problem (A.5). Notice that

$$\begin{aligned} &\operatorname{trace}\left(\left(FF^\top + D\right)^{-1}\Sigma\right) - \log\det\left(FF^\top + D\right)^{-1} \\ &= \operatorname{trace}\left((\tilde{D} - \tilde{F}\tilde{F}^\top)\Sigma\right) - \log\det\left(\tilde{D} - \tilde{F}\tilde{F}^\top\right) \end{aligned} \tag{A.8}$$

and

$$\left(FF^\top + D\right)^{-1} = \tilde{D} - \tilde{F}\tilde{F}^\top. \tag{A.9}$$

We now solve problem (A.5) with respect to $\tilde{F}$ and $\tilde{D}$:

$$\underset{\tilde{F},\tilde{D}}{\text{minimize}}\ \operatorname{trace}\left((\tilde{D} - \tilde{F}\tilde{F}^\top)\Sigma\right) - \log\det\left(\tilde{D} - \tilde{F}\tilde{F}^\top\right). \tag{A.10}$$

By taking the gradient of the objective of (A.10) with respect to $\tilde{D}$ (using the fact that $\tilde{D}$ is diagonal) and setting it to zero, we get the necessary optimality condition

$$\operatorname{diag}\Sigma = \operatorname{diag}\left(\tilde{D}^\star - \tilde{F}^\star\tilde{F}^{\star,\top}\right)^{-1}, \tag{A.11}$$

which, through (A.9), implies

$$\operatorname{diag}\Sigma = \operatorname{diag}\left(D^\star + F^\star F^{\star,\top}\right). \tag{A.12}$$

This concludes the proof and implies that any method that converges to the solution of (A.1), e.g., EM in many practical applications, actually converges to a low-rank plus diagonal model whose diagonal agrees with the diagonal of Σ. A different proof can be found in [26][Chapter 9].

We can tie this with the typical EM update for fitting a low-rank-plus-diagonal covariance model [8][Chapter 8.2], where the update for F and D using EM is

$$\begin{aligned} F_{k+1} &= C_{xs,k}C_{ss,k}^{-1} \\ D_{k+1} &= \operatorname{diag}\left(\Sigma - C_{xs,k}C_{ss,k}^{-1}C_{xs,k}\right). \end{aligned} \tag{A.13}$$

Here, we use $C_{xs,k}$ and $C_{ss,k}$ to denote the matrices C_{rf} and C_{ff}, respectively, in

[8][Chapter 8.2] at iteration k. Then (A.13) is just the update in [8][Chapter 8.2]. It holds that

$$\operatorname{diag}\Sigma_{k+1} = \operatorname{diag}\left(C_{xs,k}C_{ss,k}^{-2}C_{sx,k}\right) + \operatorname{diag}\left(\Sigma - C_{xs,k}C_{ss,k}^{-1}C_{sx,k}\right). \tag{A.14}$$

At convergence, as stated above, it must be that $\operatorname{diag}\Sigma_{k+1} \to \operatorname{diag}\Sigma$, and therefore $C_{ss,k}^{-2} \to C_{ss,k}^{-1}$ or equivalently $C_{ss,k} \to I$. It might be possible to leverage the fact that $C_{ss,k} \to I$ to speed up convergence of the EM loop. More on this in section A.5.

As an aside, if D is not restricted to be diagonal, then the necessary optimality condition is

$$\Sigma = \left(\tilde{D}^\star - \tilde{F}^\star\tilde{F}^{\star,\top}\right)^{-1}, \tag{A.15}$$

i.e., $\Sigma = D^\star + F^\star F^{\star,\top}$, which makes complete sense because of the added degrees of freedom in D.

A.3 Fitting only D

Suppose now that we fit only the diagonal matrix D:

$$\underset{D}{\text{minimize}}\,\operatorname{trace}\left(\left(FF^\top + D\right)^{-1}\Sigma\right) - \log\det\left(FF^\top + D\right)^{-1}. \tag{A.16}$$

We will take the derivative of the objective with respect to D. Let $X = D + FF^\top$. In the general case for D (not restricted to be diagonal), the differentials are

$$dX = dD, \tag{A.17}$$

$$\begin{aligned} d\,\operatorname{trace}\left((FF^\top + D)^{-1}\Sigma\right) &= \operatorname{trace}\left(d\left((FF^\top + D)^{-1}\right)\Sigma\right) \\ &= \operatorname{trace}\left(-(FF^\top + D)^{-1}\,dX\,(FF^\top + D)^{-1}\,\Sigma\right) \\ &= \operatorname{trace}\left(-(FF^\top + D)^{-1}\,\Sigma\,(FF^\top + D)^{-1}\,dD\right), \end{aligned} \tag{A.18}$$

$$d\log\det\left(FF^\top + D\right)^{-1} = -(FF^\top + D)^{-1}dD. \tag{A.19}$$

Therefore, in the general case for D, with $\Sigma^\star = FF^\top + D^\star$, the necessary optimality condition is:

$$\Sigma^{\star,-1}\,\Sigma\,\Sigma^{\star,-1} = \Sigma^{\star,-1}. \tag{A.20}$$

Multiplying by $\Sigma^\star$ left and right in both sides and simplifying, we get

$$\Sigma = \Sigma^\star. \tag{A.21}$$

Therefore for D not constrained to be diagonal, we get $\Sigma = \Sigma^\star$, which is reasonable as D has enough degrees of freedom.

For D constrained to be diagonal, the necessary optimality condition is:

$$\operatorname{diag}\Sigma^{\star,-1}\,\Sigma\,\Sigma^{\star,-1} = \operatorname{diag}\Sigma^{\star,-1}. \tag{A.22}$$

We observe that in the case of fitting only D diagonal, we are not guaranteed that the diagonal of our model will match the diagonal of Σ.

A.4 Fitting only F

Suppose now that we fit only F:

$$\underset{F}{\text{minimize}}\,\operatorname{trace}\left(\left(FF^\top + D\right)^{-1}\Sigma\right) - \log\det\left(FF^\top + D\right)^{-1}. \tag{A.23}$$

We take the derivative of the objective with respect to F. We get the following derivatives for each term in the objective:

$$\nabla_F \operatorname{trace}\left(\left(FF^\top + D\right)^{-1}\Sigma\right) = -2\left(FF^\top + D\right)^{-1}\Sigma\left(FF^\top + D\right)^{-1}F, \tag{A.24}$$

$$\nabla_F \log\det\left(FF^\top + D\right)^{-1} = -2(FF^\top + D)^{-1}F. \tag{A.25}$$

By letting the optimal point be $\Sigma^\star = F^\star F^{\star,\top} + D$, the necessary optimality condition

is

$$\Sigma^{\star,-1}\Sigma\Sigma^{\star,-1}F^{\star} = \Sigma^{\star,-1}F^{\star}. \tag{A.26}$$

In this case, we are not guaranteed that the diagonal of our model will match the diagonal of Σ.

If $F^{\star}$ is square and full-rank, then we get that the optimality condition is equivalent to

$$\Sigma^{\star,-1}\Sigma\Sigma^{\star,-1} = \Sigma^{\star,-1}, \tag{A.27}$$

i.e., we get the same optimality condition as in the case of having a fixed F and fitting a square matrix D (not constrained to be diagonal). In other words, fitting a square D given an F and fitting a square F given a D are equivalent. For $F^{\star}$ square and full-rank: $\Sigma = \Sigma^{\star}$.

A.5 Speeding-up EM

The EM update at iteration k for a factor model

$$x = Fs + \epsilon$$

is (for an arbitrary density q on s)

$$\begin{aligned}\underset{F,D}{\text{maximize}}\, \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)} &- \left(x^T D^{-1}x + \log\det D\right) \\ &- \Big(-2\,\text{trace}\left(F^T D^{-1}x\mathbb{E}_q\left[s^\top\right]\right) + \text{trace}\left((F^\top D^{-1}F)\mathbb{E}_q\left[ss^\top\right]\right)\Big).\end{aligned} \tag{A.28}$$

Although we do not include the derivation here, this can be found in [8, Chapter 8]. In the typical version of the algorithm (from hereon called *vanilla*), we use $q_k = p_k(S = s \mid X = x)$, where p_k corresponds to the distribution

$$\begin{pmatrix} S \\ X \end{pmatrix} \sim \mathcal{N}\left(0,\ \begin{pmatrix} I & F_k^\top \\ F_k & F_k F_k^\top + D_k \end{pmatrix}\right).$$

From the analysis in Section A.2, we know that no matter the sequence of densities $\{q_k\}_k$ used in each iteration, we should have that

$$C_{ss,k} =: \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{q_k}\left[ss^\top\right] \to_k I. \tag{A.29}$$

We now discuss our proposal for faster EM convergence (as given in number of iterations). Suppose we are at iteration k. Our goal is to choose q_k such that $C_{ss,k} =: \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{q_k}\left[ss^\top\right]$ is very close to I. By allowing $\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{q_k}\left[ss^\top\right]$ to converge to the identity faster, we hope to allow the EM algorithm to converge in a smaller number of iterations. We will use $q_k(s) = p_k(S = s \mid B_kX = x)$ and seek to choose B_k appropriately. The subscript k in the density denotes that we will use F_k and D_k from iteration k for the joint distribution of (S, X). Assuming an invertible B_k, it holds that $p_k(S = s \mid B_kX = x) = p_k(S = s \mid X = B_k^{-1}x)$. Therefore

$$S \mid B_kX = x \sim \mathcal{N}\left(L_kB_k^{-1}x, G_k\right), \tag{A.30}$$

where L_k and G_k are B_j and G_j, respectively, in [8, Chapter 8]. Under this model

$$\begin{aligned}\mathbb{E}_k\left[ss^\top \mid B_kX = x\right] &= G_k + G_kF_k^\top D_k^{-1}B_k^{-1}xx^\top B_k^{-\top}D_k^{-1}F_kG_k\\ \mathbb{E}_k\left[xs^\top \mid B_kX = x\right] &= xx^\top B_k^{-\top}D_k^{-1}F_kG_k\end{aligned}. \tag{A.31}$$

Therefore, the update in our case (from hereon called *q-select*) is the same as that in [8, Chapter 8], but replacing

$$\begin{aligned}C_{ss,k} &\leftarrow G_k + G_kF_k^\top D_k^{-1}B_k^{-1}\Sigma B_k^{-\top}D_k^{-1}F_kG_k\\ C_{xs,k} &\leftarrow \Sigma B_k^{-\top}D_k^{-1}F_kG_k\end{aligned}. \tag{A.32}$$

All that is left is to pick B_k at iteration k such that $C_{ss,k}$ is close to the identity. We therefore solve

$$C_{ss,k} = I. \tag{A.33}$$

By letting $A_k =: D_k^{-1}F_k$, $M_k =: A_k^\top X A_k$, $X =: B_k^{-1}\Sigma B_k^{-\top}$, we obtain

$$M = G_k^{-1}(I - G_k)G_k^{-1}, \quad T = A_k^{\dagger\top} M A_k^\dagger \tag{A.34}$$

and finally

$$B_k = \Sigma^{1/2} T^{-1/2}. \tag{A.35}$$

A.5.1 Simulations

We compare *q-select* with *vanilla* EM and *shrinkage* EM for the case of dimension $n = 50$. *Shrinkage* EM follows [8, Chapter 8], with the only exception being that it sets

$$C_{ss,k} = 0.2(L_k \Sigma L_k^\top + G_k) + 0.8I. \tag{A.36}$$

We look at the Frobenius norm of consecutive covariance estimates $\|\Sigma_{k+1} - \Sigma_k\|$ in Figure A.1 for the three methods as a function of the iteration number. In Figure A.2, we see the error $\|C_{ss,k} - I\|$ as a function of the iteration number. Finally in Figure A.3, we see the scatterplot of the diagonal of the converged estimate against the diagonal of the original covariance Σ. We note that the per iteration cost of the *q-select* is larger than the other two variants.

All methods converge to a covariance estimate that satisfies the constraint in the diagonal and *q-select* converges in fewer iterations than the other two methods. This implies that *q-select* is a promising method to speed up convergence of EM for fitting a low-rank plus diagonal covariance model.

A.6 Fitting the model using CVXPY

We show how to fit the low-rank-plus-diagonal model of (A.4) using CVXPY [95]. Because the problem is not convex in F and D, we will use a change of variables and leverage a new disciplined non-convex programming (DNLP) framework [96] to solve the problem. This framework builds on top of CVXPY [95].

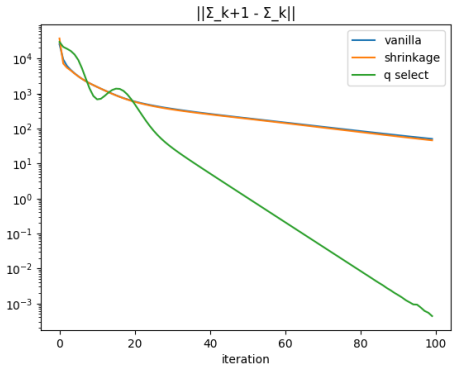


Figure A.1: The Frobenius norm of consecutive covariance estimates $\|\Sigma_{k+1} - \Sigma_k\|$ for the three methods as a function of the iteration number.

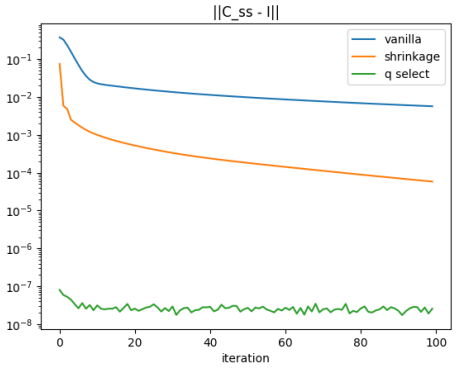


Figure A.2: The error $\|C_{ss,k} - I\|$ as a function of the iteration number for the three methods.

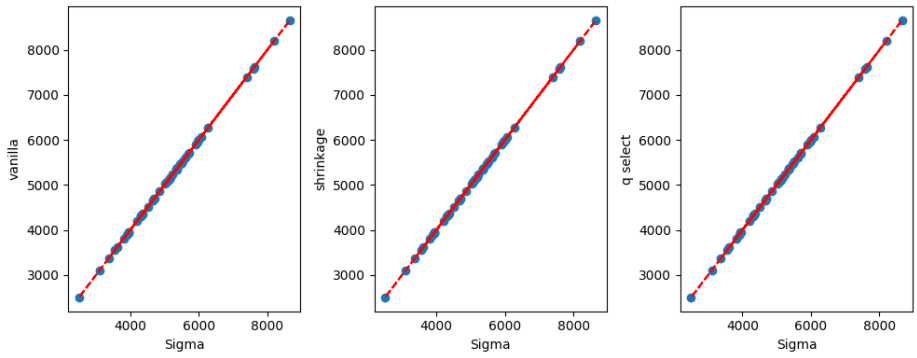


Figure A.3: Scatterplot of the diagonal of the converged estimate against the diagonal of the original covariance Σ for the three methods.

A.6.1 Method

Let the covariance model be $\hat{\Sigma} = (FF^\top + D)$. We will parametrize the optimization problem using the inverse covariance estimate $\hat{\Sigma}^{-1}$. We consider the (lower triangular) Cholesky decomposition of $\hat{\Sigma}^{-1}$:

$$\hat{\Sigma}^{-1} = LL^\top. \tag{A.37}$$

There also exist matrices E (diagonal with positive diagonal entries) and $G \in \mathbb{R}^{n\times k}$ such that

$$\hat{\Sigma}^{-1} = E - GG^\top, \tag{A.38}$$

where the mapping is

$$\begin{array}{ll} E = D^{-1} & D = E^{-1} \\ G = D^{-1}F\left(I + F^\top D^{-1} F\right)^{-1/2} & \iff \quad F = E^{-1}G\left(I - G^\top E^{-1} G\right)^{-1/2}. \end{array} \tag{A.39}$$

Let $\Sigma = RR^\top$ be the (lower-triangular) Cholesky decomposition of Σ. To maximize the Gaussian log likelihood, we solve the optimization problem

$$\begin{array}{ll} \max\limits_{L,E,G} & \sum\limits_{i=1}^{n} \log L_{ii} - \dfrac{1}{2}\|L^\top R\|_F^2 \\ \text{s.t.} & E \succ 0, \quad E \text{ diagonal} \\ & \operatorname{diag}(L) \geq 0, \quad L \text{ lower} - \text{triangular} \\ & E - GG^\top = LL^\top \end{array}, \tag{A.40}$$

The objective is the Gaussian log-likelihood. This problem is compatible with DNLP [96]. After obtaining the solution $G^\star, E^\star$, we can map it to $F^\star, D^\star$ using (A.39).

A.6.2 Observations

Another method to compute a low-rank-plus-diagonal model is expectation-maximization (EM). This method iteratively maximizes a lower-bound of the Gaussian log-likelihood.

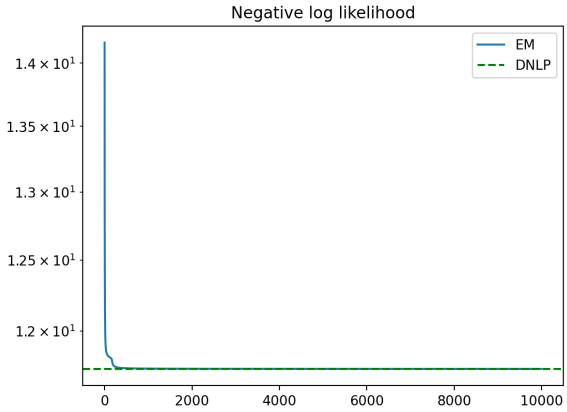


Figure A.4: Log likelihood of different methods.

Empirically, EM converges to the same covariance estimate as the DNLP solution:

$$F_{\text{EM}}F_{\text{EM}}^{\top} = F_{\text{DNLP}}F_{\text{DNLP}}^{\top}, \quad D_{\text{EM}} = D_{\text{DNLP.}} \tag{A.41}$$

Further, the achieved log-likelihood is the same for EM and DNLP. The EM and DNLP solution is different than the empirical covariance, because of the structural constraints. The empirical covariance maximizes the log likelihood under no structural constraints.

B Derivation of the EM update

We derive the update for Algorithm 2.1. Our derivation of the iterative algorithm focuses on optimizing a series of lower bounds of problem (2.11).

B.1 A lower bound of the objective in (2.11)

We fix θ, i.e., F_2, Ω, and D. Let

$$S = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} \Omega & 0 \\ 0 & I \end{bmatrix}), \quad E \sim \mathcal{N}(0, D), \tag{B.1}$$

where $S_1 \in \mathbb{R}^{n_1}, S_2 \in \mathbb{R}^{n_2}$ and $E \in \mathbb{R}^n$ are random variables. Further, we suppose S and E are independent, i.e., $S \perp E$. Similarly, we assume $S_1 \perp S_2$. We define

$$X = \begin{bmatrix} F_1 & F_2 \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} + E. \tag{B.2}$$

It follows that

$$\begin{bmatrix} S \\ X \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \tilde{\Omega} & \tilde{\Omega}\tilde{F}^\top \\ \tilde{F}\tilde{\Omega} & \tilde{F}\tilde{\Omega}\tilde{F}^\top + D \end{bmatrix}\right). \tag{B.3}$$

Marginally X is distributed according to $\mathcal{N}(0, \tilde{F}\tilde{\Omega}\tilde{F}^\top + D)$ with density

$$p_\theta(x) = \int_s p_\theta(x, s) ds, \tag{B.4}$$

where $p_\theta(x, s)$ is the joint density of (X, S) for the given θ, i.e., F_2, Ω, and D, evaluated at $X = x$ and $S = s$.

We now also fix x. Let $q(s;x)$ be a probability density function in s, parametrized by x. Then

$$\log p_\theta(x) = \log \int_s p_\theta(x,s)ds \tag{B.5}$$

$$= \log \int_s q(s;x)\frac{p_\theta(x,s)}{q(s;x)}ds = \log \mathbb{E}_{q(s;x)}\frac{p_\theta(x,s)}{q(s;x)} \geq \mathbb{E}_{q(s;x)} \log \frac{p_\theta(x,s)}{q(s;x)}, \tag{B.6}$$

where the inequality follows from Jensen's inequality for the concave function log. Because eq. (B.6) holds for all x, it is true that

$$\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)} \log p_\theta(x) \geq \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{q(s;x)} \log \frac{p_\theta(x,s)}{q(s;x)}. \tag{B.7}$$

This implies that

$$\underset{\theta}{\text{maximize}}\, \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)} \log p_\theta(x) \geq \underset{\theta}{\text{maximize}}\, \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{q(s;x)} \log \frac{p_\theta(x,s)}{q(s;x)}, \tag{B.8}$$

where the optimal point of the right-hand side is the same as that of

$$\underset{\theta}{\text{maximize}}\, \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{q(s;x)} \log p_\theta(x,s). \tag{B.9}$$

Notice that we can write

$$p_\theta(x,s) = p_\theta(X=x, S=s) = p_\theta(E = x - \tilde{F}s, S=s) = p_\theta(E = x-\tilde{F}s)p_\theta(S=s), \tag{B.10}$$

because $S \perp E$. Therefore, up to an additive constant,

$$2\log p_\theta(x,s) = -(x-\tilde{F}s)^\top D^{-1}(x-\tilde{F}s) - \log\det D - s^\top\tilde{\Omega}^{-1}s - \log\det\tilde{\Omega}. \tag{B.11}$$

By expanding eq. (B.11) and using

$$\tilde{\Omega}^{-1} = \begin{bmatrix} \Omega^{-1} & 0 \\ 0 & I \end{bmatrix} \tag{B.12}$$

we get

$$\begin{aligned} &-2\log p_\theta(x,s) = \\ &x^\top D^{-1}x - 2\begin{bmatrix} s_1^\top & s_2^\top \end{bmatrix}\begin{bmatrix} F_1^\top \\ F_2^\top \end{bmatrix} D^{-1}x + s_1^\top F_1^\top D^{-1}F_1 s_1 + 2s_2^\top F_2^\top D^{-1}F_1 s_1 + s_2^\top F_2^\top D^{-1}F_2 s_2 \\ &+ s_1^\top \Omega^{-1} s_1 + s_2^\top s_2 + \log\det D + \log\det\Omega. \end{aligned} \tag{B.13}$$

For any density $q(s;x)$, ignoring all terms that do not depend on F_2, Ω or D, using the trace trick (trace of a scalar equals the scalar), and the cyclic property of the trace, we get

$$\begin{aligned} &\mathbb{E}_{q(s;x)}\left[-2\log p_\theta(x,s)\right] = \\ &\quad \operatorname{trace}(D^{-1}xx^\top) + \operatorname{trace}(D^{-1}F_1\mathbb{E}_{q(s;x)}\left[s_1 s_1^\top\right]F_1^\top) - 2\operatorname{trace}(D^{-1}\mathbb{E}_{q(s;x)}\left[xs_1^\top\right]F_1^\top) \\ &\quad - 2\operatorname{trace}(D^{-1}\mathbb{E}_{q(s;x)}\left[xs_2^\top\right]F_2^\top) + 2\operatorname{trace}(D^{-1}F_1\mathbb{E}_{q(s;x)}\left[s_1 s_2^\top\right]F_2^\top) \\ &\quad + \operatorname{trace}(D^{-1}F_2\mathbb{E}_{q(s;x)}\left[s_2 s_2^\top\right]F_2^\top) \\ &\quad + \operatorname{trace}(\Omega^{-1}\mathbb{E}_{q(s;x)}\left[s_1 s_1^\top\right]) \\ &\quad + \log\det D + \log\det\Omega. \end{aligned} \tag{B.14}$$

At this point, we have an explicit expression for the objective of problem (B.9).

B.2 The EM update

The EM algorithm is an iterative algorithm that solves a sequence of problems of the form (B.9). At each iteration the density $q(s;x)$ is updated. Assuming that at

iteration k we are provided with θ_k, the EM algorithm computes θ_{k+1} as

$$\begin{aligned}\theta_{k+1} &= \arg\max_{\theta} \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{q_k(s;x)} \log p_\theta(x,s) \\ &= \arg\min_{\theta} \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{q_k(s;x)} \left[-2\log p_\theta(x,s)\right],\end{aligned} \tag{B.15}$$

where $q_k(s;x) = p_{\theta_k}(S = s \mid X = x)$. We therefore need to compute

- $\mathbb{E}_{q_k(s;x)}\left[s_1\right] =: \mathbb{E}\left[S_1 \mid X = x;\theta_k\right]$
- $\mathbb{E}_{q_k(s;x)}\left[s_2\right] =: \mathbb{E}\left[S_2 \mid X = x;\theta_k\right]$,
- $\mathbb{E}_{q_k(s;x)}\left[s_1 s_2^\top\right] =: \mathbb{E}\left[S_1 S_2^\top \mid X = x;\theta_k\right]$,
- $\mathbb{E}_{q_k(s;x)}\left[s_2 s_2^\top\right] =: \mathbb{E}\left[S_2 S_2^\top \mid X = x;\theta_k\right]$,
- $\mathbb{E}_{q_k(s;x)}\left[s_1 s_1^\top\right] =: \mathbb{E}\left[S_1 S_1^\top \mid X = x;\theta_k\right]$.

This is easy to do under our Gaussian assumptions. Note that

$$\begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \mid X = x;\theta_k \sim \mathcal{N}(L_k x, G_k). \tag{B.16}$$

We can compute L_k and G_k using $p_{\theta_k}(S = s \mid X = x)$. Because, $S \mid X = x$ is Gaussian:

- $\mathbb{E}\left[S \mid X = x;\theta_k\right] = \arg\max_s \log p_{\theta_k}(S = s \mid X = x) = \arg\max_s \log p_{\theta_k}(x,s)$. By taking the derivative of (B.11) with respect to s and setting it to zero, we get the necessary optimality condition

$$-\tilde{F}_k^\top D_k^{-1} x + \tilde{F}_k^\top D_k^{-1}\tilde{F}_k s + \tilde{\Omega}_k^{-1} s = 0 \Leftrightarrow s = \left(\tilde{F}_k^\top D_k^{-1}\tilde{F}_k + \tilde{\Omega}_k^{-1}\right)^{-1} \tilde{F}_k^\top D_k^{-1} x. \tag{B.17}$$

  This implies that

$$\mathbb{E}\left[S \mid X = x;\theta_k\right] = \left(\tilde{F}_k^\top D_k^{-1}\tilde{F}_k + \tilde{\Omega}_k^{-1}\right)^{-1} \tilde{F}_k^\top D_k^{-1} x. \tag{B.18}$$

- The inverse of the conditional covariance is given by the Hessian of $-\log p_{\theta_k}(S = s \mid X = x)$ with respect to s, i.e.,

$$G_k^{-1} = \tilde{F}_k^\top D_k^{-1} \tilde{F}_k + \tilde{\Omega}_k^{-1}. \tag{B.19}$$

We combine the two results and get that

$$L_k = G_k \tilde{F}_k^\top D_k^{-1}, \quad G_k^{-1} = \tilde{F}_k^\top D_k^{-1} \tilde{F}_k + \tilde{\Omega}_k^{-1}. \tag{B.20}$$

Further by the definition of (conditional) covariance,

$$\mathbb{E}\left[SS^\top \mid X = x; \theta_k\right] = G_k + \mathbb{E}\left[S \mid X = x; \theta_k\right] \mathbb{E}\left[S \mid X = x; \theta_k\right]^\top, \tag{B.21}$$

which can equivalently be written as

$$\begin{bmatrix} \mathbb{E}\left[S_1 S_1^\top \mid X = x; \theta_k\right] & \mathbb{E}\left[S_1 S_2^\top \mid X = x; \theta_k\right] \\ \mathbb{E}\left[S_2 S_1^\top \mid X = x; \theta_k\right] & \mathbb{E}\left[S_2 S_2^\top \mid X = x; \theta_k\right] \end{bmatrix} = G_k + L_k x x^\top L_k^\top. \tag{B.22}$$

Notice that we can select the appropriate block from $\mathbb{E}\left[SS^\top \mid X = x; \theta_k\right]$ by multiplying left and right with the appropriate (block-selector) matrices. For example:

- $$\mathbb{E}\left[S_1 S_1^\top \mid X = x; \theta_k\right] = B_{l,11}(G_k + L_k x x^\top L_k^\top) B_{r,11}, \tag{B.23}$$

  where

$$B_{l,11} = \begin{bmatrix} I_{n_1} & 0_{n_1 \times n_2} \end{bmatrix}, \quad B_{r,11} = \begin{bmatrix} I_{n_1} \\ 0_{n_2 \times n_1} \end{bmatrix}. \tag{B.24}$$

  and I_n is the identity matrix of dimension n.

- $$\mathbb{E}\left[S_2 S_2^\top \mid X = x; \theta_k\right] = B_{l,22}(G_k + L_k x x^\top L_k^\top) B_{r,22}, \tag{B.25}$$

where

$$B_{l,22} = \begin{bmatrix} 0_{n_2 \times n_1} & I_{n_2} \end{bmatrix}, \quad B_{r,22} = \begin{bmatrix} 0_{n_1 \times n_2} \\ I_{n_2} \end{bmatrix}. \tag{B.26}$$

- $$\mathbb{E}\left[S_1 S_2^\top \mid X = x; \theta_k\right] = B_{l,12}(G_k + L_k x x^\top L_k^\top) B_{r,12}, \tag{B.27}$$

  where

$$B_{l,12} = \begin{bmatrix} I_{n_1} & 0_{n_1 \times n_2} \end{bmatrix}, \quad B_{r,12} = \begin{bmatrix} 0_{n_1 \times n_2} \\ I_{n_2} \end{bmatrix}. \tag{B.28}$$

Similarly

- $$\mathbb{E}\left[S_1 \mid X = x; \theta_k\right] = B_1 L_k x, \tag{B.29}$$

  where $B_1 = \begin{bmatrix} I_{n_1} & 0_{n_1 \times n_2} \end{bmatrix}$.

- $$\mathbb{E}\left[S_2 \mid X = x; \theta_k\right] = B_2 L_k x, \tag{B.30}$$

  where $B_2 = \begin{bmatrix} 0_{n_2 \times n_1} & I_{n_2} \end{bmatrix}$.

By combining the above results with (B.14), we get that

$$\begin{aligned}
&\mathbb{E}_{q_k(s;x)}\left[-2 \log p_\theta(x, s)\right] = \\
&\quad \operatorname{trace}(D^{-1} x x^\top) + \operatorname{trace}(D^{-1} F_1 B_{l,11} (G_k + L_k x x^\top L_k^\top) B_{r,11} F_1^\top) \\
&\quad - 2 \operatorname{trace}(D^{-1} x x^\top L_k^\top B_1^\top F_1^\top) \\
&\quad - 2 \operatorname{trace}(D^{-1} x x^\top L_k^\top B_2^\top F_2^\top) \\
&\quad + 2 \operatorname{trace}(D^{-1} F_1 B_{l,12} \left(G_k + L_k x x^\top L_k^\top\right) B_{r,12} F_2^\top) \\
&\quad + \operatorname{trace}(D^{-1} F_2 B_{l,22} \left(G_k + L_k x x^\top L_k^\top\right) B_{r,22} F_2^\top) \\
&\quad + \operatorname{trace}(\Omega^{-1} B_{l,11} \left(G_k + L_k x x^\top L_k^\top\right) B_{r,11}) \\
&\quad + \log\det D + \log\det \Omega
\end{aligned} \tag{B.31}$$

and therefore

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{q_k(s;x)}\left[-2\log p_\theta(x,s)\right] = \\ &\operatorname{trace}(D^{-1}\Sigma) + \operatorname{trace}(D^{-1}F_1B_{l,11}(G_k + L_k\Sigma L_k^\top)B_{r,11}F_1^\top) \\ &- 2\operatorname{trace}(D^{-1}\Sigma L_k^\top B_1^\top F_1^\top) \\ &- 2\operatorname{trace}(D^{-1}\Sigma L_k^\top B_2^\top F_2^\top) \\ &+ 2\operatorname{trace}(D^{-1}F_1B_{l,12}\left(G_k + L_k\Sigma L_k^\top\right)B_{r,12}F_2^\top) \\ &+ \operatorname{trace}(D^{-1}F_2B_{l,22}\left(G_k + L_k\Sigma L_k^\top\right)B_{r,22}F_2^\top) \\ &+ \operatorname{trace}(\Omega^{-1}B_{l,11}\left(G_k + L_k\Sigma L_k^\top\right)B_{r,11}) \\ &+ \log\det D + \log\det\Omega.\end{aligned} \tag{B.32}$$

At this point, we note that for any distribution $\mathbb{P}$ with zero mean and covariance Σ

$$\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{q_k(s;x)}\left[-2\log p_\theta(x,s)\right] = \mathbb{E}_{x\sim\mathcal{P}}\mathbb{E}_{q_k(s;x)}\left[-2\log p_\theta(x,s)\right]. \tag{B.33}$$

We will use this result later to show that the EM update is the same no matter the distribution of x, as long as this distribution is zero-mean with covariance Σ.

We now define

$$\begin{aligned}C_{ss}^k &=: G_k + L_k\Sigma L_k^\top \\ &= \begin{bmatrix} C_{ss,11}^k & C_{ss,12}^k \\ C_{ss,12}^{k,\top} & C_{ss,22}^k \end{bmatrix} = \begin{bmatrix} B_{l,11}(G_k + L_k\Sigma L_k^\top)B_{r,11} & B_{l,12}(G_k + L_k\Sigma L_k^\top)B_{r,12} \\ \left(B_{l,12}(G_k + L_k\Sigma L_k^\top)B_{r,12}\right)^\top & B_{l,22}(G_k + L_k\Sigma L_k^\top)B_{r,22} \end{bmatrix}\end{aligned} \tag{B.34}$$

and partition

$$L_k = \begin{bmatrix} L_1^k \\ L_2^k \end{bmatrix}, \tag{B.35}$$

where $L_1^k \in \mathbb{R}^{n_1\times n}$ and $L_k^2 \in \mathbb{R}^{n_2\times n}$. We further define

- $$\tilde{\Sigma}_k =: \Sigma + F_1C_{ss,11}^kF_1^\top - 2\Sigma L_1^{k,\top}F_1^\top \tag{B.36}$$

•

$$C_{xs}^k =: \Sigma L_2^{k,\top} - F_1 C_{ss,12}^k \tag{B.37}$$

With these definitions

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{q_k(s;x)}\left[-2\log p_\theta(x,s)\right] = \\ &\operatorname{trace}\left(D^{-1}\left(\tilde{\Sigma}_k - 2C_{xs}^k F_2^\top + F_2 C_{ss,22}^k F_2^\top\right)\right) \\ &+ \operatorname{trace}(\Omega^{-1} C_{ss,11}^k) \\ &+ \log\det\Omega + \log\det D.\end{aligned} \tag{B.38}$$

B.3 Explicit equations for the EM update

Using (B.38), we are now ready to compute the EM update

$$\theta_{k+1} = \arg\min_\theta \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{q_k(s;x)}\left[-2\log p_\theta(x,s)\right] \tag{B.39}$$

under the constraint that D is diagonal. Notice that the problem is separable in (D, F_2) and Ω.

We do the easy part first and find the optimal value for Ω. To do this we take the derivative of $\operatorname{trace}(\Omega^{-1}C_{ss,11}) + \log\det\Omega$ with respect to Ω^{-1} and set it to zero. This gives:

$$\Omega^{\star,\top} = C_{ss,11}^{k,\top} \Rightarrow \Omega^\star = C_{ss,11}^k. \tag{B.40}$$

We now turn to the sub-problem that depends on D, F:

$$\operatorname*{minimize}_D \operatorname*{minimize}_{F_2} \operatorname{trace}\left(D^{-1}\left(\tilde{\Sigma}_k - 2C_{xs}^k F_2^\top + F_2 C_{ss,22}^k F_2^\top\right)\right) + \log\det D. \tag{B.41}$$

We fix D. The optimal F_2 for this D can be found by taking the derivative of the objective with respect to F_2 and setting it to zero. This gives

$$D^{-1}(-2C_{xs}^k + 2C_{ss,22}^k F_2^\star(D)) = 0 \Leftrightarrow F_2^\star(D) = C_{xs}^k C_{ss,22}^{k,-1}. \tag{B.42}$$

It only remains to minimize over the diagonal D. We plug $F_2^\star(D)$ in the objective of (B.41), which takes care of the internal minimization with respect to F. To obtain the optimal diagonal D, we solve

$$\underset{D}{\text{minimize}} \operatorname{trace}\left(D^{-1}\left(\tilde{\Sigma}_k - C_{xs}^k C_{ss,22}^{k,-1} C_{xs}^{k,\top}\right)\right) + \log\det D. \tag{B.43}$$

Let $D_{ii} = 1/\psi_i$. Then, problem (B.43) is equivalent to

$$\underset{\psi_1,\ldots,\psi_n}{\text{minimize}} \sum_{i=1}^n \psi_i \left(\tilde{\Sigma}_k - C_{xs}^k C_{ss,22}^{k,-1} C_{xs}^{k,\top}\right)_{ii} - \log\psi_i. \tag{B.44}$$

The last problem is separable in the ψ_i's. By taking the derivatives with respect to the ψ_i's and setting them to zero, we get the solution

$$\frac{1}{\psi_i^\star} = \left(\tilde{\Sigma}_k - C_{xs}^k C_{ss,22}^{k,-1} C_{xs}^{k,\top}\right)_{ii}. \tag{B.45}$$

Therefore,

$$D^\star = \operatorname{diag}\left(\tilde{\Sigma} - C_{xs} C_{ss,22}^{-1} C_{xs}^\top\right). \tag{B.46}$$

To summarize, the EM update is

$$\begin{aligned} \Omega &\leftarrow C_{ss,11}^k \\ F_2 &\leftarrow C_{xs}^k C_{ss,22}^{k,-1} \\ D &\leftarrow \operatorname{diag}\left(\tilde{\Sigma}_k - C_{xs}^k C_{ss,22}^{k,-1} C_{xs}^{k,\top}\right). \end{aligned} \tag{B.47}$$

B.3.1 Interpretation as maximum likelihood estimation

By (B.33), for any distribution $\mathbb{P}$ with zero mean and covariance Σ, the EM update satisfies

$$\arg\min_\theta \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)} \mathbb{E}_{q_k(s;x)}\left[-2\log p_\theta(x,s)\right] = \arg\min_\theta \mathbb{E}_{x\sim\mathbb{P}} \mathbb{E}_{q_k(s;x)}\left[-2\log p_\theta(x,s)\right]. \tag{B.48}$$

This implies that minimizing the KL divergence between $\mathbb{P}$ and $\mathcal{N}(0, \tilde{F}\tilde{\Omega}\tilde{F}^\top + D)$ with EM is equivalent to minimizing the KL divergence between $\mathcal{N}(0, \Sigma)$ and $\mathcal{N}(0, \tilde{F}\tilde{\Omega}\tilde{F}^\top + D)$ with EM. The update only depends on the covariance Σ, which is common for $\mathbb{P}$ and $\mathcal{N}(0, \Sigma)$. This justifies calling this method maximum likelihood covariance estimation, as we are fitting a Gaussian to empirical data.

B.4 Evidence lower bound (ELBO) for EM

The EM algorithm maximizes a sequence of lower bounds, as seen by

$$\underset{\theta}{\text{maximize}}\, \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)} \log p_\theta(x) \geq \underset{\theta}{\text{maximize}}\, \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)} \mathbb{E}_{q(s;x)} \log \frac{p_\theta(x,s)}{q(s;x)} \tag{B.49}$$

for any $q(x;s)$. If we set $q(s;x)$ to be the conditional distribution $p_\theta(s \mid x)$, then

$$\log \frac{p_\theta(x,s)}{q(s;x)} = \log p_\theta(x) \Rightarrow \mathbb{E}_{q(s;x)} \log \frac{p_\theta(x,s)}{q(s;x)} = \log p_\theta(x). \tag{B.50}$$

Let $J(\theta) = \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)} \log p_\theta(x)$. Then, by eq. (B.7)

$$J(\theta_{k+1}) \geq \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)} \mathbb{E}_{q_k(s;x)} \log \frac{p_{\theta_{k+1}}(x,s)}{q_k(s;x)} \tag{B.51}$$

and by (B.15)

$$\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)} \mathbb{E}_{q_k(s;x)} \log \frac{p_{\theta_{k+1}}(x,s)}{q_k(s;x)} \geq \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)} \mathbb{E}_{q_k(s;x)} \log \frac{p_{\theta_k}(x,s)}{q_k(s;x)}. \tag{B.52}$$

By (B.50)

$$\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)} \mathbb{E}_{q_k(s;x)} \log \frac{p_{\theta_k}(x,s)}{q_k(s;x)} = \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)} \log p_{\theta_k}(x) = J(\theta_k). \tag{B.53}$$

Combining the above results

$$J(\theta_{k+1}) \geq J(\theta_k) \tag{B.54}$$

for the EM update. In other words, EM improves the expected log-likelihood at every iteration.

C The EM update with missing returns data

We now derive the EM update in the case of missing returns data. We first describe the set-up and then derive the EM update.

C.1 The set-up

We derive the EM algorithm in the case of missing returns data. We consider a time window of length T. Let X_τ be the returns random vector at time $\tau = 1, \ldots, T$. We assume that at time τ we only observe the returns X_τ on a subset of assets indexed by the set $\mathcal{O}_\tau \subseteq \{1, \ldots, n\}$. Let $\mathcal{M}_\tau = \{1, \ldots, n\} \setminus \mathcal{O}_\tau$ be the set of assets with missing returns at time τ. We denote the observed portion of the return at τ as $X_{\tau,\text{obs}}$ (with value $x_{\tau,\text{obs}}$) and the missing portion as $X_{\tau,\text{mis}}$. We define the permutation matrix $\Pi_\tau \in \mathbb{R}^{n \times n}$ that satisfies

$$\begin{bmatrix} X_{\tau,\text{mis}} \\ X_{\tau,\text{obs}} \end{bmatrix} = \Pi_\tau X_\tau. \tag{C.1}$$

By direct analogy to problem (2.11), we wish to solve

$$\underset{\theta}{\text{maximize}} \sum_{\tau=1}^{T} w_\tau \log p_{\theta,\text{obs},\tau}(x_{\tau,\text{obs}}), \tag{C.2}$$

where $p_\theta(x)$ is the probability density function of the Gaussian $\mathcal{N}(0, \tilde{F}\tilde{\Omega}\tilde{F}^\top + D)$,

$$\tilde{F} = \begin{bmatrix} F_1 & F_2 \end{bmatrix}, \qquad \tilde{\Omega} = \begin{bmatrix} \Omega & 0 \\ 0 & I \end{bmatrix}, \tag{C.3}$$

$p_{\theta,\text{obs},\tau}(\cdot)$ is the marginal density of the observed returns at time τ, $\theta =: (F_2, \Omega, D)$, and w_τ are non-negative weights that sum to 1.

It is true that

$$\begin{aligned} \log p_{\theta,\text{obs},\tau}(x_{\tau,\text{obs}}) = \log \int p_\theta(x_\tau, s_\tau) dx_{\tau,\text{mis}} ds_\tau = \\ \log \int q(s_\tau, x_{\tau,\text{mis}}; x_{\tau,\text{obs}}) \frac{p_\theta(x_\tau, s_\tau)}{q(s_\tau, x_{\tau,\text{mis}}; x_{\tau,\text{obs}})} dx_{\tau,\text{mis}} ds_\tau \geq \\ \mathbb{E}_{q(s_\tau, x_{\tau,\text{mis}}; x_{\tau,\text{obs}})} \log \frac{p_\theta(x_\tau, s_\tau)}{q(s_\tau, x_{\tau,\text{mis}}; x_{\tau,\text{obs}})}, \end{aligned} \tag{C.4}$$

where $x_\tau = \Pi_\tau^\top \begin{bmatrix} x_{\tau,\text{mis}} \\ x_{\tau,\text{obs}} \end{bmatrix}$, and $q(s_\tau, x_{\tau,\text{mis}}; x_{\tau,\text{obs}})$ is any density over the latent variables s_τ and the missing returns $x_{\tau,\text{mis}}$ parameterized by the observed returns $x_{\tau,\text{obs}}$. The inequality follows from Jensen's inequality.

Instead of solving (C.2) directly, we can maximize the lower bound

$$\underset{\theta}{\text{maximize}} \sum_{\tau=1}^{T} w_\tau \mathbb{E}_{q(s_\tau, x_{\tau,\text{mis}}; x_{\tau,\text{obs}})} \log \frac{p_\theta(x_\tau, s_\tau)}{q(s_\tau, x_{\tau,\text{mis}}; x_{\tau,\text{obs}})} \tag{C.5}$$

that has the same solution as

$$\underset{\theta}{\text{minimize}} \sum_{\tau=1}^{T} w_\tau \mathbb{E}_{q(s_\tau, x_{\tau,\text{mis}}; x_{\tau,\text{obs}})} \left[-2 \log p_\theta(x_\tau, s_\tau) \right]. \tag{C.6}$$

Note that similar to Chapter B,

$$\begin{aligned}
&\mathbb{E}_{q_\tau}\left[-2\log p_\theta(x_\tau, s_\tau)\right] = \\
&\quad \operatorname{trace}(D^{-1}x_\tau x_\tau^\top) + \operatorname{trace}(D^{-1}F_1\mathbb{E}_{q_\tau}\left[s_{\tau,1}s_{\tau,1}^\top\right]F_1^\top) - 2\operatorname{trace}(D^{-1}\mathbb{E}_{q_\tau}\left[x_\tau s_{\tau,1}^\top\right]F_1^\top) \\
&\quad - 2\operatorname{trace}(D^{-1}\mathbb{E}_{q_\tau}\left[x_\tau s_{\tau,2}^\top\right]F_2^\top) + 2\operatorname{trace}(D^{-1}F_1\mathbb{E}_{q_\tau}\left[s_{\tau,1}s_{\tau,2}^\top\right]F_2^\top) \\
&\quad + \operatorname{trace}(D^{-1}F_2\mathbb{E}_{q_\tau}\left[s_{\tau,2}s_{\tau,2}^\top\right]F_2^\top) \\
&\quad + \operatorname{trace}(\Omega^{-1}\mathbb{E}_{q_\tau}\left[s_{\tau,1}s_{\tau,1}^\top\right]) \\
&\quad + \log\det D + \log\det\Omega,
\end{aligned} \tag{C.7}$$

where we abbreviate $q(s_\tau, x_{\tau,\text{mis}}; x_{\tau,\text{obs}})$ as q_τ for simplicity.

EM solves (C.6) iteratively. Assuming we are at iteration k, with θ_k, we set

$$q(s_\tau, x_{\tau,\text{mis}}; x_{\tau,\text{obs}}) = p_{\theta_k}(S_\tau = s_\tau, X_{\tau,\text{mis}} = x_{\tau,\text{mis}} \mid X_{\tau,\text{obs}} = x_{\tau,\text{obs}}). \tag{C.8}$$

We remind the reader that for arbitrary θ, the random variables satisfy

$$\begin{aligned}
X_{\tau,\text{obs}} &= \tilde{F}_{\tau,\text{obs}}S_\tau + E_{\tau,\text{obs}} \\
X_{\tau,\text{mis}} &= \tilde{F}_{\tau,\text{mis}}S_\tau + E_{\tau,\text{mis}}
\end{aligned}, \tag{C.9}$$

where

$$S_\tau = \begin{bmatrix} S_{\tau,1} \\ S_{\tau,2} \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} \Omega & 0 \\ 0 & I \end{bmatrix}), \quad E_\tau \sim \mathcal{N}(0, D), \tag{C.10}$$

and S_τ is independent of E_τ. Also, to make the notation clear: $\tilde{F}_{\tau,\text{obs}}$ refers to the rows of $\tilde{F}_\tau$ indexed by $\mathcal{O}_\tau$, and similarly for the other quantities indexed by obs or mis.

These relations allows us to compute the conditional distribution in (C.8) and the necessary statistical moments, as we show next.

C.2 Computing the statistical moments

Similar to Chapter B, the conditional distribution of S_τ given the observed returns $X_{\tau,\text{obs}} = x_{\tau,\text{obs}}$ at iteration k (i.e., with $\tilde{F}_k, \Omega_k, D_k$) is

$$S_\tau \mid X_{\tau,\text{obs}} = x_{\tau,\text{obs}}; \theta_k \sim \mathcal{N}(L_{k,\tau} x_{\tau,\text{obs}}, G_{k,\tau}), \tag{C.11}$$

where

$$L_{k,\tau} = G_{k,\tau} \tilde{F}_{k,\tau,\text{obs}}^\top D_{k,\tau,\text{obs}}^{-1}, \quad G_{k,\tau}^{-1} = \tilde{F}_{k,\tau,\text{obs}}^\top D_{k,\tau,\text{obs}}^{-1} \tilde{F}_{k,\tau,\text{obs}} + \tilde{\Omega}_k^{-1}. \tag{C.12}$$

To make the notation clear, $\tilde{F}_{k,\tau,\text{obs}}$ is the sub-matrix of $\tilde{F}_k$ that contains the rows indexed by $\mathcal{O}_\tau$ and $D_{k,\tau,\text{obs}}$ is the sub-matrix of D_k that contains the rows and columns indexed by $\mathcal{O}_\tau$.

We also define

$$\hat{s}_\tau =: L_{k,\tau} x_{\tau,\text{obs}}. \tag{C.13}$$

We can now compute the necessary statistical moments at iteration k. First,

$$\mathbb{E}\left[S_\tau S_\tau^\top \mid X_{\tau,\text{obs}} = x_{\tau,\text{obs}}; \theta_k\right] = G_{k,\tau} + \hat{s}_\tau \hat{s}_\tau^\top. \tag{C.14}$$

Second,

$$\mathbb{E}\left[X_{\tau,\text{obs}} S_\tau^\top \mid X_{\tau,\text{obs}} = x_{\tau,\text{obs}}; \theta_k\right] = x_{\tau,\text{obs}} \hat{s}_\tau^\top. \tag{C.15}$$

Third,

$$\begin{aligned} \mathbb{E}\left[X_{\tau,\text{mis}} S_\tau^\top \mid X_{\tau,\text{obs}} = x_{\tau,\text{obs}}; \theta_k\right] &= \tilde{F}_{k,\tau,\text{mis}} \mathbb{E}\left[S_\tau S_\tau^\top \mid X_{\tau,\text{obs}} = x_{\tau,\text{obs}}; \theta_k\right] \\ &= \tilde{F}_{k,\tau,\text{mis}} (\hat{s}_\tau \hat{s}_\tau^\top + G_{k,\tau}) = \hat{x}_{\tau,\text{mis}} \hat{s}_\tau^\top + \tilde{F}_{k,\tau,\text{mis}} G_{k,\tau}, \end{aligned} \tag{C.16}$$

where $\hat{x}_{\tau,\text{mis}} = \tilde{F}_{k,\tau,\text{mis}} \hat{s}_\tau$ are the imputed missing values. Equations (C.15)-(C.16)

imply that

$$\mathbb{E}\left[X_\tau S_\tau^\top \mid X_{\tau,\text{obs}} = x_{\tau,\text{obs}}; \theta_k\right] = \underbrace{\Pi_\tau^\top \begin{bmatrix} \hat{x}_{\tau,\text{mis}} \\ x_{\tau,\text{obs}} \end{bmatrix}}_{\hat{x}_\tau} \hat{s}_\tau^\top + \Pi_\tau^\top \begin{bmatrix} \tilde{F}_{k,\tau,\text{mis}} G_{k,\tau} \\ 0 \end{bmatrix} \tag{C.17}$$

and

$$\mathbb{E}\left[X_\tau X_\tau^\top \mid X_{\tau,\text{obs}} = x_{\tau,\text{obs}}; \theta_k\right] = \hat{x}_\tau \hat{x}_\tau^\top + \Pi_\tau^\top \begin{bmatrix} \tilde{F}_{k,\tau,\text{mis}} G_{k,\tau} \tilde{F}_{k,\tau,\text{mis}}^\top + D_{k,\tau,\text{mis}} & 0 \\ 0 & 0 \end{bmatrix} \Pi_\tau. \tag{C.18}$$

C.3 Obtaining the update equations

We are at iteration k, with θ_k. Let

- $C_{xx}^k =: \sum_{\tau=1}^T w_\tau \mathbb{E}\left[X_\tau X_\tau^\top \mid X_{\tau,\text{obs}} = x_{\tau,\text{obs}}; \theta_k\right]$,
- $W_{xs}^k =: \sum_{\tau=1}^T w_\tau \mathbb{E}\left[X_\tau S_\tau^\top \mid X_{\tau,\text{obs}} = x_{\tau,\text{obs}}; \theta_k\right]$,
- $C_{ss}^k =: \sum_{\tau=1}^T w_\tau \mathbb{E}\left[S_\tau S_\tau^\top \mid X_{\tau,\text{obs}} = x_{\tau,\text{obs}}; \theta_k\right]$.

According to EM, we solve problem (C.6) using (C.8) to get θ_{k+1}. With our new definitions, we can write problem (C.6) as

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \text{trace}\Big(D^{-1}\Big[C_{xx}^k + F_1 C_{ss,11}^k F_1^\top - 2W_{xs,1}^k F_1^\top - 2W_{xs,2}^k F_2^\top \\ & \qquad + 2F_1 C_{ss,12}^k F_2^\top + F_2 C_{ss,22}^k F_2^\top\Big]\Big) \\ & + \log\det D \\ & + \text{trace}\big(\Omega^{-1} C_{ss,11}^k\big) + \log\det\Omega. \end{aligned} \tag{C.19}$$

We further define

- $\tilde{\Sigma}_k =: C_{xx}^k + F_1 C_{ss,11}^k F_1^\top - 2W_{xs,1}^k F_1^\top$,

- $C_{xs}^k =: W_{xs,2}^k - F_1 C_{ss,12}^k$.

With these definitions, (C.19) becomes

$$\begin{aligned} \underset{\theta}{\text{minimize}} \, & \text{trace}\left(D^{-1}\left(\tilde{\Sigma}_k - 2C_{xs}^k F_2^\top + F_2 C_{ss,22}^k F_2^\top\right)\right) \\ & + \log\det D \\ & + \text{trace}(\Omega^{-1} C_{ss,11}^k) + \log\det\Omega. \end{aligned} \tag{C.20}$$

We have already solved this problem in Chapter B. Therefore, the EM update with missing returns data is

$$\begin{aligned} \Omega &\leftarrow C_{ss,11}^k \\ F_2 &\leftarrow C_{xs}^k C_{ss,22}^{k,-1} \\ D &\leftarrow \text{diag}\left(\tilde{\Sigma}_k - C_{xs}^k C_{ss,22}^{k,-1} C_{xs}^{k,\top}\right) \end{aligned} \quad . \tag{C.21}$$

Bibliography

[1] H. Markowitz, "Portfolio Selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[2] R. C. Grinold and R. N. Kahn, *Active Portfolio Management*. McGraw Hill New York, 2000.

[3] A. J. McNeil, R. Frey, and P. Embrechts, *Quantitative Risk Management: Concepts, Techniques and Tools-Revised Edition*. Princeton University Press, 2015.

[4] W. F. Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *The Journal of Finance*, vol. 19, no. 3, pp. 425–442, 1964.

[5] S. Boyd, K. Johansson, R. Kahn, P. Schiele, and T. Schmelzer, "Markowitz Portfolio Construction at Seventy," *Journal of Portfolio Management*, vol. 50, no. 8, pp. 117–160, 2024.

[6] C. Stărică and C. Granger, "Nonstationarities in Stock Returns," *Review of Economics and Statistics*, vol. 87, no. 3, pp. 503–522, 2005.

[7] J. Fan, Y. Liao, and H. Liu, "An Overview of the Estimation of Large Covariance and Precision Matrices," *The Econometrics Journal*, vol. 19, no. 1, pp. C1–C32, 2016.

[8] K. Johansson, M. G. Ogut, M. Pelger, T. Schmelzer, and S. Boyd, "A Simple Method for Predicting Covariance Matrices of Financial Returns," *Foundations and Trends® in Econometrics*, vol. 12, no. 4, pp. 324–407, 2023.

[9] MSCI Inc., *MSCI Barra Risk Models*, `https://app2.msci.com/products/analytics/models/`.

[10] J. Menchero, D Orr, and J. Wang, "The Barra US Equity Model (USE4), Methodology Notes," *MSCI Barra*, 2011.

[11] R. F. Engle, "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica: Journal of the Econometric Society*, pp. 987–1007, 1982.

[12] E. F. Fama and K. R. French, "The Cross-Section of Expected Stock Returns," *The Journal of Finance*, vol. 47, no. 2, pp. 427–465, 1992.

[13] E. F. Fama and K. R. French, "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, vol. 33, no. 1, pp. 3–56, 1993.

[14] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[15] S. Pafka and I. Kondor, "Noisy Covariance Matrices and Portfolio Optimization II," *Physica A: Statistical Mechanics and its Applications*, vol. 319, pp. 487–494, 2003.

[16] O. Ledoit and M. Wolf, "Honey, I Shrunk the Sample Covariance Matrix," *The Journal of Portfolio Management*, vol. 30, no. 4, pp. 110–119, 2004.

[17] R. JP Morgan, *Riskmetrics–Technical Document*, 1996.

[18] S. Barratt and S. Boyd, "Covariance Prediction via Convex Optimization," *Optimization and Engineering*, vol. 24, no. 3, pp. 2045–2078, 2023.

[19] R. Engle, "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models," *Journal of Business & Economic Statistics*, vol. 20, no. 3, pp. 339–350, 2002.

[20] F. J. Fabozzi, *Handbook of Finance, Investment Management and Financial Management*. Wiley, 2008, vol. 2.

[21] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd. Springer, 2009.

[22] J. Bai and S. Ng, "Large Dimensional Factor Analysis," *Foundations and Trends® in Econometrics*, vol. 3, no. 2, pp. 89–163, 2008.

[23] K. Yeon and M. Anitescu, "Beyond Low Rank: Fast Low-Rank + Diagonal Decomposition with a Spectral Approach," *arXiv preprint arXiv:2512.17120*, 2025.

[24] D. B. Rubin and D. T. Thayer, "EM Algorithms for ML Factor Analysis," *Psychometrika*, vol. 47, no. 1, pp. 69–76, 1982.

[25] A.-K. Seghouane, "An Iterative Projections Algorithm for ML Factor Analysis," in *IEEE Workshop on Machine Learning for Signal Processing*, 2008.

[26] K. V. Mardia, J. T. Kent, and C. C. Taylor, *Multivariate Analysis*. John Wiley & Sons, 2024.

[27] A. Spector, R. F. Barber, T. Hastie, R. N. Kahn, and E. Candès, "The Mosaic Permutation Test: An Exact and Nonparametric Goodness-of-Fit Test for Factor Models," *arXiv preprint arXiv:2404.15017*, 2024.

[28] W. Lee, R. Brown, and H. de Silva, "Narrative Factors and Risk Models," *Available at SSRN 5271271*, 2025.

[29] E. Candès, T. Hastie, K. Hogan, R. N. Kahn, R. Luo, and A. Spector, "Thematic Investing: A Risk-Based Perspective," *Financial Analysts Journal*, vol. 81, no. 4, pp. 103–120, 2025.

[30] G. Rapakoulias and P. Tsiotras, "Discrete-Time Maximum Likelihood Neural Distribution Steering," in *IEEE Conference on Decision and Control (CDC)*, 2024.

[31] I. M. Balci and E. Bakolas, "Covariance Steering of Discrete-Time Stochastic Linear Systems based on Wasserstein Distance Terminal Cost," *IEEE Control Systems Letters*, vol. 5, no. 6, 2020.

[32] G. Rapakoulias and P. Tsiotras, "Discrete-Time Optimal Covariance Steering via Semidefinite Programming," in *IEEE Conference on Decision and Control (CDC)*, 2023.

[33] G. Mathew and I. Mezić, "Metrics for Ergodicity and Design of Ergodic Dynamics for Multi-Agent Systems," *Physica D: Nonlinear Phenomena*, vol. 240, no. 4-5, 2011.

[34] L. Dressel and M. J. Kochenderfer, "Tutorial on the Generation of Ergodic Trajectories with Projection-based Gradient Descent," *IET Cyber-Physical Systems: Theory & Applications*, vol. 4, no. 2, 2019.

[35] L. M. Miller and T. D. Murphey, "Trajectory Optimization for Continuous Ergodic Exploration," in *American Control Conference*, 2013.

[36] M. Araya, O. Buffet, V. Thomas, and F. Charpillet, "A POMDP Extension with Belief-Dependent Rewards," *Advances in Neural Information Processing Systems*, 2010.

[37] K. Nadjahi, A. Durmus, L. Chizat, S. Kolouri, S. Shahrampour, and U. Simsekli, "Statistical and Topological Properties of Sliced Probability Divergences," *Advances in Neural Information Processing Systems*, 2020.

[38] C. R. Givens and R. M. Shortt, "A Class of Wasserstein Metrics for Probability Distributions.," *Michigan Mathematical Journal*, vol. 31, no. 2, 1984.

[39] S. Grossi, M. Letizia, and R. Torre, "Refereeing the Referees: Evaluating Two-Sample Tests for Validating Generators in Precision Sciences," *Machine Learning: Science and Technology*, vol. 6, 2025.

[40] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, 2007.

[41] S. Kolouri, G. K. Rohde, and H. Hoffmann, "Sliced Wasserstein Distance for Learning Gaussian Mixture Models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[42] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde, "Generalized Sliced Wasserstein Distances," *Advances in Neural Information Processing Systems*, 2019.

[43] H. Su and H. Zhang, "Distances and Kernels based on Cumulative Distribution Functions," in *Emerging Trends in Image Processing, Computer Vision and Pattern Recognition*, Elsevier, 2015.

[44] A. E. Tzikas, L. Fiechtner, A. Jamgochian, and M. J. Kochenderfer, "Distributionally Robust Control with Constraints on Linear Unidimensional Projections," *arXiv preprint arXiv:2508.07121*, 2025.

[45] I. Deshpande, Z. Zhang, and A. G. Schwing, "Generative Modeling using the Sliced Wasserstein Distance," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[46] J. Rabin, G. Peyré, J. Delon, and M. Bernot, "Wasserstein Barycenter and its Application to Texture Mixing," in *International Conference on Scale Space & Variational Methods in Computer Vision*, 2011.

[47] T. Lin, Z. Zheng, E. Chen, M. Cuturi, and M. I. Jordan, "On Projection Robust Optimal Transport: Sample Complexity and Model Misspecification," in *International Conference on Artificial Intelligence and Statistics*, 2021.

[48] V. Titouan, R. Flamary, N. Courty, R. Tavenard, and L. Chapel, "Sliced Gromov-Wasserstein," in *Advances in Neural Information Processing Systems*, 2019.

[49] I. Kim, S. Balakrishnan, and L. Wasserman, "Robust Multivariate Nonparametric Tests via Projection Averaging," *The Annals of Statistics*, vol. 48, no. 6, 2020.

[50] Z. Li and Y. Zhang, "On a Projective Ensemble Approach to Two Sample Test for Equality of Distributions," in *International Conference on Machine Learning*, 2020.

[51] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola, "A Kernel Two-Sample Test," *Journal of Machine Learning Research*, vol. 13, pp. 723–773, 2012.

[52] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet, "Hilbert Space Embeddings and Metrics on Probability Measures," *Journal of Machine Learning Research*, vol. 11, pp. 1517–1561, 2010.

[53] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, "Kernel Mean Embedding of Distributions: A Review and Beyond," *Foundations and Trends in Machine Learning*, vol. 10, no. 1–2, pp. 1–141, 2017.

[54] B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet, "Universality, Characteristic Kernels and RKHS Embedding of Measures," *Journal of Machine Learning Research*, vol. 12, pp. 2389–2410, 2011.

[55] W. Kramer, *Probability Measure*. Wiley-Interscience, 1995.

[56] P. Billingsley, *Probability and Measure*. John Wiley & Sons, 2017.

[57] M. J. Kochenderfer, T. A. Wheeler, and K. H. Wray, *Algorithms for Decision Making*. MIT Press, 2022.

[58] A. E. Tzikas, D. Knowles, G. X. Gao, and M. J. Kochenderfer, "Multirobot Navigation using Partially Observable Markov Decision Processes with Belief-based Rewards," *Journal of Aerospace Information Systems*, vol. 20, no. 8, pp. 437–446, 2023.

[59] A. E. Tzikas, J. Park, M. J. Kochenderfer, and R. E. Allen, "Distributed Online Planning for Min-Max Problems in Networked Markov Games," *IEEE Robotics and Automation Letters*, vol. 9, no. 7, pp. 6656–6663, 2024.

[60] Y. Bengio, I. Goodfellow, and A. Courville, *Deep Learning*. MIT Press, 2017.

[61] H. W. Kuhn, "The Hungarian Method for the Assignment Problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1–2, pp. 83–97, 1955.

[62] G. Hinton, N. Srivastava, and K. Swersky, *Lecture 6e: RMSProp: Divide the Gradient by a Running Average of its Recent Magnitude*, Neural Networks for Machine Learning, Coursera, 2012.

[63] J. Duchi, E. Hazan, and Y. Singer, "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.

[64] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[65] I. Loshchilov and F. Hutter, *Decoupled Weight Decay Regularization*, 2019. arXiv: 1711.05101.

[66] A. N. Angelopoulos and S. Bates, "Conformal Prediction: A Gentle Introduction," *Foundations and Trends in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023.

[67] A. N. Angelopoulos, R. F. Barber, and S. Bates, "Theoretical Foundations of Conformal Prediction," *arXiv preprint arXiv:2411.11824*, 2024.

[68] I. Gibbs, J. J. Cherian, and E. J. Candès, "Conformal Prediction with Conditional Guarantees," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 87, no. 4, pp. 1100–1126, 2025.

[69] Y. Qin, R. Wang, A. J. Vakharia, Y. Chen, and M. M. Seref, "The Newsvendor Problem: Review and Directions for Future Research," *European Journal of Operational Research*, vol. 213, no. 2, pp. 361–374, 2011.

[70] T.-M. Choi, D. Li, and H. Yan, "Mean–Variance Analysis for the Newsvendor Problem," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 38, no. 5, pp. 1169–1180, 2008.

[71] G. E. Monahan, N. C. Petruzzi, and W. Zhao, "The Dynamic Pricing Problem from a Newsvendor's Perspective," *Manufacturing & Service Operations Management*, vol. 6, no. 1, pp. 73–91, 2004.

[72] H. Bavafa, C. M. Leys, L. Örmeci, and S. Savin, "Managing Portfolio of Elective Surgical Procedures: A Multidimensional Inverse Newsvendor Problem," *Operations Research*, vol. 67, no. 6, pp. 1543–1563, 2019.

[73] B. Marchi, S. Zanoni, and M. Pasetti, "Multi-Period Newsvendor Problem for the Management of Battery Energy Storage Systems in Support of Distributed Generation," *Energies*, vol. 12, no. 23, p. 4598, 2019.

[74] M. Schwenzer, M. Ay, T. Bergs, and D. Abel, "Review on Model Predictive Control: An Engineering Perspective," *The International Journal of Advanced Manufacturing Technology*, vol. 117, no. 5, pp. 1327–1349, 2021.

[75] N. C. Petruzzi and M. Dada, "Pricing and the Newsvendor Problem: A Review with Extensions," *Operations Research*, vol. 47, no. 2, pp. 183–194, 1999.

[76] K. Talluri and G. Van Ryzin, "Revenue Management under a General Discrete Choice Model of Consumer Behavior," *Management Science*, vol. 50, no. 1, pp. 15–33, 2004.

[77] R. Levi, G. Perakis, and J. Uichanco, "The Data-Driven Newsvendor Problem: New Bounds and Insights," *Operations Research*, vol. 63, no. 6, pp. 1294–1306, 2015.

[78] C.-T. See and M. Sim, "Robust Approximation to Multiperiod Inventory Management," *Operations Research*, vol. 58, no. 3, pp. 583–594, 2010.

[79] F. Caro and J. Gallien, "Inventory Management of a Fast-Fashion Retail Network," *Operations Research*, vol. 58, no. 2, pp. 257–273, 2010.

[80] A. V. Den Boer, "Dynamic Pricing and Learning: Historical Origins, Current Research, and New Directions," *Surveys in Operations Research and Management Science*, vol. 20, no. 1, pp. 1–18, 2015.

[81] G. Kim, K. Wu, and E. Huang, "Optimal Inventory Control in a Multi-Period Newsvendor Problem with Non-Stationary Demand," *Advanced Engineering Informatics*, vol. 29, no. 1, pp. 139–145, 2015.

[82] H. Wang, B. Chen, and H. Yan, "Optimal Inventory Decisions in a Multiperiod Newsvendor Problem with Partially Observed Markovian Supply Capacities," *European Journal of Operational Research*, vol. 202, no. 2, pp. 502–517, 2010.

[83] D. B. Khang and O. Fujiwara, "Optimality of Myopic Ordering Policies for Inventory Model with Stochastic Supply," *Operations Research*, vol. 48, no. 1, pp. 181–184, 2000.

[84] M. Densing, "Dispatch Planning using Newsvendor Dual Problems and Occupation Times: Application to Hydropower," *European Journal of Operational Research*, vol. 228, no. 2, pp. 321–330, 2013.

[85] R. Wang, X. Yan, and C. Zhu, "Solving a Distribution-Free Multi-Period Newsvendor Problem with Advance Purchase Discount via an Online Ordering Solution," *SAGE Open*, vol. 13, no. 2, 2023.

[86] H. Sarimveis, P. Patrinos, C. D. Tarantilis, and C. T. Kiranoudis, "Dynamic Modeling and Control of Supply Chain Systems: A Review," *Computers & Operations Research*, vol. 35, no. 11, pp. 3530–3561, 2008.

[87] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2009.

[88] K. Border, "Differentiating an Integral: Leibniz' Rule," *Caltech Division of the Humanities and Social Sciences*, 2016.

[89] M. J. Kochenderfer and T. A. Wheeler, *Algorithms for Optimization*. MIT Press, 2019.

[90] M. Udell and S. Boyd, "Bounding Duality Gap for Separable Problems with Linear Constraints," *Computational Optimization and Applications*, vol. 64, no. 2, pp. 355–378, 2016.

[91] M. Udell and S. Boyd, "Maximizing a Sum of Sigmoids," *Optimization and engineering*, pp. 1–25, 2013.

[92] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*. MIT Press, 2016.

[93] N. Balakrishnan, N. Johnson, and S. Kotz, *Continuous Univariate Distributions*. John Wiley & Sons, Incorporated, 2016.

[94] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2006.

[95] S. Diamond and S. Boyd, "Cvxpy: A Python-embedded Modeling Language for Convex Optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.

[96] D. Cederberg, W. Zhang, P. Nobel, and S. Boyd, "Disciplined Nonlinear Programming," 2025.