Title
Robust Optimization of Generalized Eigenvalue Problem for Positive Semi-definite Matrices

Author
Wang, Jiaming

Publication Date
2023

Peer reviewed|Thesis/dissertation

Robust Optimization of Generalized Eigenvalue Problem for Positive Semi-definite Matrices

by

Jiaming Wang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Applied Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Ming Gu, Chair
Professor Jon Wilkening
Professor Per-Olof Persson

Spring 2023

Robust Optimization of Generalized Eigenvalue Problem for Positive Semi-definite Matrices

Abstract

Robust Optimization of Generalized Eigenvalue Problem for Positive Semi-definite Matrices

by

Jiaming Wang

Doctor of Philosophy in Applied Mathematics

University of California, Berkeley

Professor Ming Gu, Chair

In this paper, we propose novel algorithms for solving the worst-case robust optimization of the generalized eigenvalue problem with positive semi-definite constraint, a highly non-convex problem that has eluded traditional optimization solvers. We consider the rank-one case and the general-rank case saparately. For the rank-one case, we first present a relaxation of the KKT system, transforming the problem into a more tractable nonlinear system of equations. We then prove the tightness of the relaxation at the optimal point and introduce two algorithms for solving the relaxed system of equations. Our approach is the first to guarantee finding the global optimal solution for the problem at hand. For the general-rank case, we first solve the KKT system with one variable fixed. We then describe an algorithm searching for the variable. We then showed that our algorithm converges to a stationary point of the problem. We showcase the potential applications of our algorithms in robust adaptive beamforming and semi-supervised Fisher discriminant analysis. This work contributes significantly to the field by providing a globally optimal solution to a highly non-convex problem with broad applicability in various disciplines.

Contents

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Ming Gu, whose unwavering support, guidance, and encouragement throughout my PhD journey have been truly invaluable. Prof. Gu's expertise, patience, and insightful feedback have not only shaped my academic growth but have also inspired me to strive for excellence in my research endeavors.

In addition to my advisor, I would like to express my sincere appreciation to Professors Zeyu Zheng and Professor Xin Guo for their assistance and support in various aspects of my academic journey. Although their contributions may not be directly related to my dissertation, their guidance and encouragement have undoubtedly had a positive impact on my overall PhD experience.

I would like to extend my heartfelt thanks to my fellow graduate students, including Haixiang Zhang, Haoting Zhang, and Jingxu Xu for their camaraderie, intellectual exchanges, and moral support during the challenging moments of my PhD journey. Their friendship, shared experiences, and collaborative spirit have made the academic process more enjoyable and rewarding, and I am grateful to have had the opportunity to learn from and alongside such talented individuals.

I am grateful to the administrative and technical staff at UC Berkeley, particularly Vicky Lee, Jon Phillips, and Clay Calder, for their assistance in facilitating a conducive research environment. Their dedication, professionalism, and attention to detail have ensured that I had the necessary resources and support to complete my dissertation successfully. Their hard work behind the scenes has truly made a difference in my academic experience.

Lastly, I would like to express my profound appreciation to my family, especially my parents, for their unconditional love, understanding, and unwavering belief in my abilities. Their constant encouragement and emotional support have been the backbone of my academic journey, and I dedicate this accomplishment to them.

In conclusion, I am indebted to all those who have played a part in my PhD journey, directly or indirectly, for their kindness, expertise, and encouragement. I am truly fortunate to have had such an exceptional network of support throughout this challenging yet fulfilling chapter of my academic career.

Chapter 1

Introduction and Background

1.1 Generalized Eigenvalue Optimization

Generalized Eigenvalue Problem

Eigenvalue and generalized eigenvalue problems are fundamental problems in numerical analysis, with broad applications across various disciplines, including computer science, economics, engineering, physics, and statistics. In an eigenvalue problem, we solve the leading eigenvectors of a matrix, which serve as essential indicators of the most significant and informative directions inherent to that matrix. On the other hand, in a generalized eigenvalue problem, the directions of the leading generalized eigenvectors are influenced by a pair of matrices A and B. More specifically, the generalized eigenvalue problem[6] is to find generalized eigenvalues λ and generalized eigenvectors $x \neq 0$ such that

$$A\,x = \lambda\,B\,x \tag{1.1}$$

or in the general-rank case,

$$AX = BX\Lambda \tag{1.2}$$

Note that the eigenvalue problem is a special case of generalized eigenvalue problem when $B = I$.

When the matrices A and B are symmetric positive semi-definite, the generalized eigenvalue problems (1.1),(1.2) are closely related to the following generalized eigenvalue optimization:

$$\max_{x \neq 0} \quad \frac{x^T\,A\,x}{x^T\,B\,x} \tag{1.3}$$

or in the general-rank case,

$$\begin{aligned} \max_{X \neq 0} \quad & \mathrm{Tr}(X^T A X) \\ \text{s.t.} \quad & X^T B X = I \end{aligned} \tag{1.4}$$

This is also called Rayleigh-quotient optimization. The optimal values and solution vectors of (1.4) are the leading generalized eigenvalues and generalized eigenvectors of (1.1),(1.2) respectively. Details about the relationship between generalized eigenvalue problem and generalized eigenvalue optimization can be found in section 2.

Adaptive Beamforming

One of the applications of generalized eigenvalue problem is adaptive beamforming. Adaptive beamforming is an advanced signal processing technique used primarily in sensor arrays for directional signal transmission or reception. This method dynamically adjusts the phases and amplitudes of the signals in the array in such a way that it effectively 'shapes' the beam pattern, enhancing the signal quality in certain directions while suppressing interference from others. The directionality of the beam can be altered based on the desired application, whether it's radar, sonar, wireless communication, or medical imaging. It relies on the concept of spatial filtering, which uses the spatial signatures of signals, such as their direction of arrival, to distinguish between the signals of interest and the unwanted noise or interference. Thus, adaptive beamforming enhances system performance by improving signal clarity, reducing noise, and increasing signal-to-noise ratio.Adaptive beamforming has found numerous applications in radar[32],[22],[26],[3], sonar[18],[2], seismology[24], microphone array[43] speech processing, and wireless communications[31],[40]. Mathematically, adaptive beamforming solves the following optimization problem:

$$\max_{w \neq 0} \quad \frac{w^H R_s w}{w^H R_{i+n} w} \tag{1.5}$$

where $w \in \mathbf{C}^{n \times 1}$ is the complex vector of beamformer weights, and R_s and R_{i+n} are the signal and interference-plus-noise covariance matrices respectively. The definitions and properties of these two matrices can be found in section 5. The problem (1.5) is a complex version of the generalized eigenvalue optimization problem (1.3).

Fisher Discriminant Analysis

Another application of generalized eigenvalue problem is Fisher discriminant analysis. Fisher Discriminant Analysis (FDA), also known as Linear Discriminant Analysis (LDA), is a statistical technique used in pattern recognition, machine learning,

and data classification. The method reduces the dimensionality of a dataset while preserving as much of the class discriminatory information as possible[39][28]. The FDA operates by projecting high-dimensional data onto a line and then performs the classification in this lower-dimensional space. The line is chosen to maximize the between-class scatter and minimize the within-class scatter, which enhances class separation. This means that similar data points (belonging to the same class) are grouped closer together, while different data points (belonging to different classes) are situated further apart. Therefore, FDA is a powerful tool for feature extraction, improving the efficiency and accuracy of classification tasks, and is widely used in fields such as image recognition and bioinformatics. LDA can easily be extended from binary cases to multi-class scenarios[39]. Mathematically, Fisher discriminant analysis solves the following optimization problem:

$$\max_{w \neq 0} \quad \frac{w^T S_b w}{w^T S w} \tag{1.6}$$

or in the general-rank case:

$$\max_{W \neq 0} \quad \mathrm{Tr}((W^T S W)^{-1} W^T S_b W) \tag{1.7}$$

where $w \in \mathbf{R}^{n \times 1}$ and $W \in \mathbf{R}^{n \times k}$ are the projection directions and projection subspaces, respectively. S_b is the between-class scatters, and S is the within-class scatters. The definitions and properties of these two matrices can be found in section 6. The problems (1.6) and (1.7) are generalized eigenvalue optimization problems (1.3),(1.4).

1.2 Robust Optimization

General Robust Optimization

New challenges arise when there is data or parameter uncertainty. In recent years, researchers proposed the concept of robust optimization to solve the general problem of data uncertainty in optimization. The main development phase of the robust counterpart methodology in convex optimization was initialized and significantly driven by the work of Ben-Tal and Nemirovski [11], [9], [10] and also independently by the work of El-Ghaoui and Lebret [20]. These approaches are based on convex optimization techniques [14] and make intensive use of the concept of duality in convex programming, which helps us to transform an important class of min-max optimization problems into tractable convex optimization problems. Here, a commonly proposed assumption is that the uncertainty set is ellipsoidal, which means

that for a vector a, if we believe that the data is uncertain, we can assume that the true vector lies in the following set:

$$U = \{\hat{a} + Ru \mid \|u\|_2 \leq 1\}$$

For example, a linear program (LP)

$$\begin{aligned} \min_x \quad & c^T x \\ \text{s.t.} \quad & ax \leq b \end{aligned} \tag{1.8}$$

with uncertain data a can be formulated as

$$\begin{aligned} \min_x \max_{a \in U} \quad & c^T x \\ \text{s.t.} \quad & ax \leq b \end{aligned} \tag{1.9}$$

This is called the robust counterpart of problem (1.8), which is a min-max optimization problem. This can be transformed into the following second order cone program (SOCP)[20]:

$$\begin{aligned} \min_x \quad & c^T x \\ \text{s.t.} \quad & \hat{a}x + \|R^T x\|_2 \leq b \end{aligned} \tag{1.10}$$

Similarly, an uncertain SOCP can – at least if the uncertainty set has a particularly structured ellipsoidal format – again be written as an SOCP. Note that the field of research addressing robust convex optimization problems has expanded during the last years and is still in progress, as reported in [7], [12]. For an extensive overview on robust optimization from the convex perspective, we refer to the text book by Ben-Tal, El-Ghaoui, and Nemirovski [8].

Robust Generalized Eigenvalue Optimization

The idea of robust optimization motivates the following robust counterpart of generalized eigenvalue problem:

$$\max_{x \neq 0} \min_{A \in S_A, B \in S_B} \quad \frac{x^T A\, x}{x^T B\, x} \tag{1.11}$$

or in the general-rank case,

$$\max_{X \neq 0} \min_{A \in S_A, B \in S_B} \quad \mathrm{Tr}((X^T B X)^{-1} X^T A X) \tag{1.12}$$

where S_A, S_B are the uncertainty sets to be determined.
The idea of robust optimization was first applied to the generalized eigenvalue problem (1.11) in 2003 in the context of robust adaptive beamforming [41]. In [41], the uncertainty set S_A, S_B are chosen to be:

$$S_A = \{A + \Delta A \mid \|\Delta A\| \leq \epsilon_A\}$$

$$S_B = \{B + \Delta B \mid \|\Delta B\| \leq \epsilon_B\}$$

In this case, the robust counterpart of the problem (1.11) is:

$$\max_{x \neq 0} \min_{\|\Delta A\| \leq \epsilon_A, \|\Delta B\| \leq \epsilon_B} \frac{x^T(A + \Delta A)x}{x^T(B + \Delta B)x} \tag{1.13}$$

This max-min optimization problem can be transformed into a generalized eigenvalue optimization problem[41]:

$$\max_{x \neq 0} \frac{x^T(A - \epsilon_A I)x}{x^T(B + \epsilon_B I)x} \tag{1.14}$$

Details about the derivation and discussion can be found in section 3.
Although the robust generalized eigenvalue problem formulated in [41] has a straightforward closed-form solution, it is excessively conservative, as the worst-case matrix A, B could be indefinite or even negative definite. Consequently, less conservative approaches were introduced in [15],[42], incorporating an additional positive semi-definite (PSD) constraint on the worst-case signal covariance matrix. This is achieved by introducing a matrix decomposition of the positive semi-definite matrix and putting the error term into both of the matrices obtained from the decomposition. In [15],[42], the uncertainty set are chosen to be:

$$\tilde{S}_A = \{(\tilde{A} + \Delta\tilde{A})(\tilde{A} + \Delta\tilde{A})^T \mid \|\Delta\tilde{A}\| \leq \epsilon_A\} \tag{1.15}$$

$$S_B = \{B + \Delta B \mid \|\Delta B\| \leq \epsilon_B\}$$

where $A = \tilde{A}^T\tilde{A}$
In this case, the robust counterpart of the problem (1.11) is:

$$\max_{x} \min_{\|\Delta\tilde{A}\| \leq \epsilon_A, \|\Delta B\| \leq \epsilon_B} \frac{\|(\tilde{A} + \Delta\tilde{A})x\|_2^2}{x^T(B + \Delta B)x} \tag{1.16}$$

This max-min optimization problem is highly non-convex and details about the derivation and discussion can be found in section 3.

Robust Adaptive Beamforming

It is widely recognized that adaptive beamforming methods' performance significantly declines when the desired signal is included in the training data, even with minor discrepancies in the knowledge of the desired signal covariance matrix[41],[22],[17],[34]. Such mismatches between the assumed and actual source covariance matrices can arise due to factors like antenna element displacement, changing environments, or imperfections in the propagation medium, among others. The primary objective of any robust adaptive beamforming (RAB) approach is to ensure resilience against these types of mismatches.

Numerous methods for robust adaptive beamforming have been proposed. One large class of methods are designed for signal look direction mismatches. One of the most popular approach in this class is the linearly constrained minimum variance (LCMV) beamformer[38]:

$$\begin{aligned} \min_{w\neq 0} \quad & w^H R_{i+n} w \\ \text{s.t.} \quad & w^H a_i a_i^H w = 1, i = 1, \ldots, k \end{aligned} \tag{1.17}$$

where a_i are presumed signal candidates. Other methods in this class include signal blocking-based algorithms[22] and Bayesian beamformer[29], etc. These methods perform well against signal look direction mismatches but are less effective against other mismatches, including calibration errors, unknown sensor coupling, wavefront mismodeling, distortions, source spreading, and both coherent and incoherent local scattering[1], indicating limitations in their applicability.

Another large class of methods are designed to be robust against more general types of mismatches. One of the most popular approach in this class is diagonal loading[5]:

$$\max_{w\neq 0} \quad \frac{w^H R_s w}{w^H (R_{i+n} + \lambda I) w} \tag{1.18}$$

We simply add a multiple of identity matrix to B and solve for the adaptive beamforming problem with the modified matrices. Diagonal loading is known to be very sensitive to the choice of the paramter λ. Other methods in this class include the eigenspace-based beamformer[17], covariance matrix tapper (CMT) approach[27], and the aforementioned robust optimization approach[41], etc. As discussed in the previous section, robust optimization approach is to solve the following robust counterpart:

$$\max_{w\neq 0} \min_{R_s \in S_1, R_{i+n} \in S_2} \quad \frac{w^H R_s w}{w^H R_{i+n} w} \tag{1.19}$$

where S_1, S_2 are uncertainty sets. For an extensive overview on robust adaptive beamforming, we refer to the book by Ayman Elnashar [19].

Semi-supervised Linear Discriminant Analysis

LDA has been shown to perform well compared to other supervised dimensionality reduction methods in experiments[33]. However, LDA requires instance-label pairs, which can be restrictive for large training datasets. In recent years, semi-supervised methods have been developed to use unlabeled data to support classification or regression tasks when labeled data is limited. When we have label information, like in classification tasks, LDA can perform much better than PCA[36]. However, when there aren't enough training samples compared to the number of dimensions, we might not accurately estimate each group's mean vector and covariance matrix. In this situation, we can't guarantee good results on test samples. A possible solution is learning from both labeled and unlabeled data, which is also called semi-supervised learning. Several approaches have been proposed, most of which are based on transductive learning[16][33][44]. This approach makes sense because, in real life, we often have only some labeled data and a large amount of unlabeled data. In [16], a semi-supervised dimensionality reduction algorithm called Semi-supervised Discriminant Analysis (SDA) is proposed. SDA aims to find a projection which respects the discriminant structure inferred from the labeled data points, as well as the intrinsic geometrical structure inferred from both labeled and unlabeled data points. In our work, we connect robust generalized eigenvalue problem (1.11),(1.12) and linear discriminant analysis and provide a novel algorithm for semi-supervised linear discriminant analysis. Details can be found in section 6.

1.3 Related Algorithms

To solve the generalized eigenvalue problem (1.1)

$$A\,x = \lambda\,B\,x$$

LAPACK [4] provides fast and accurate implementations. The corresponding generalized eigenvalue optimization (1.3) can then be solved. In the context of adaptive beamforming, solving for (1.3) directly is known as the Minimum Variance Distortionless Response (MVDR) algorithm.
As discussed earlier, the robust generalized eigenvalue problem (1.16)

$$\max_{x} \min_{\|\Delta\tilde{A}\|\leq\epsilon_A,\|\Delta B\|\leq\epsilon_B} \frac{\|(\tilde{A}+\Delta\tilde{A})x\|_2^2}{x^T(B+\Delta B)x}$$

is highly non convex and eludes traditional optimization solvers. Researchers have been solving this problem by developing algorithms to solve its convex approximations.

In [15], researhcers first transformed the problem (1.16) to:

$$\begin{aligned} \min_{X\neq 0} \quad & \mathrm{Tr}((B+\epsilon_B I)X) \\ \text{s.t.} \quad & \mathrm{Tr}(\tilde{A}^T\tilde{A}X) - \epsilon_A^2\mathrm{Tr}(X) - 1 \geq 2\epsilon_A\sqrt{\mathrm{Tr}(X)} \\ & X \geq 0 \\ & \mathrm{rank}(X) = 1 \end{aligned} \tag{1.20}$$

where $X = xx^H$ is a matrix, and this is a non-convex problem due to the rank constraint. Researchers then used a series of Semi-definite program (SDP) to approximate the non-convex problem.
In [42], researchers derived closed form solutions to two convex modifications of the problem (1.16):

$$x_1^{opt} = \mathcal{P}((B+\epsilon_B I)^{-1}(\tilde{A}^T\tilde{A} - \epsilon_A^2 I)) \tag{1.21}$$

$$x_2^{opt} = \mathcal{P}((B+\epsilon_B I)^{-1}(\tilde{A}^T\tilde{A} - 2\sqrt{\lambda_{max}(\tilde{A}^T\tilde{A})}\eta I + \epsilon_A^2 I)) \tag{1.22}$$

where $\mathcal{P}(\cdot)$ denotes the leading eigenvector operator. The benefit of approximating with this closed form solution is that the efficiency of the method is as good as the original adaptive beamforming problem and thus the robust formulation does not incur additional computational cost.
In [30], researchers rewrote the problem as the minimization of a one-dimensional optimal value function and then further converted it to a convex SDP problem. They showed that their algorithm converges to the global optimal under certain conditions.
Recently, [23] proposed a new SOCP based algorithm to approximate the solution to the robust generalized eigenvalue problem, which avoids solving semi-definite program and is therefore faster.
The main drawback of the methods in [15],[42] is that they only provide a suboptimal solution, potentially leaving a significant gap to the global optimal solution. For instance, the method in [15] iteratively finds a suboptimal solution, but there is no guarantee of convergence. A closed-form approximate suboptimal solution is proposed in [42]; however, this solution might also be far from the global optimal one. These limitations prompt the exploration of new, efficient strategies to solve the aforementioned non-convex problem in a globally optimal manner.

1.4 Structure of the Thesis

The rest of the paper is organized as follows. Generalized eigenvalue problem is introduced in Chapter 2. The rank-one case robust generalized eigenvalue problem is

discussed in Chapter 3. The general-rank case robust generalized eigenvalue problem is discussed in Chapter 4. Applications and numerical experiments are shown in Chapter 5 and 6 for robust adaptive beamforming and semi-supervised linear discriminant analysis, respectively.

Chapter 2

Generalized Eigenvalue Problem

In this chapter, we first introduce eigenvalue problem and generalized eigenvalue problem. And then we formulate them as different forms of optimization problems. Finally, we introduce several applications of generalized eigenvalue problems. This chapter is mostly from a tutorial paper[21] of this topic.

2.1 Eigenvalue Problem

The eigenvalue problem[25] of a matrix A is to find eigenvalues λ and eigenvectors x such that

$$Ax = \lambda x \tag{2.1}$$

Or in matrix form, it is

$$AX = X\Lambda \tag{2.2}$$

where X consists of columns of eigenvectors and Λ is a diagonal matrix of eigenvalues When A is symmetric, X is orthogonal matrix and all diagonals of Λ are real numbers. Furthermore, when A is positive definite, diagonals of Λ are positive real numbers.

2.2 Generalized Eigenvalue Problem

The generalized eigenvalue problem[6] of two matrices A and B is to find generalized eigenvalues λ and generalized eigenvectors x such that

$$Ax = \lambda Bx \tag{2.3}$$

Or in matrix form, it is

$$AX = BX\Lambda \tag{2.4}$$

where X consists of columns of eigenvectors and Λ is a diagonal matrix of eigenvalues The generalized eigenvalue problem is usually denoted by a pair of matrices (A, B), which is also called 'pencil'[6]. It is obvious that eigenvalue problem is a special case of generalized eigenvalue problem when $B = I$.

2.3 Eigenvalue Optimization

Symmetric eigenvalue problem is closely related to some specific optimization problems. [21] summarizes five forms of eigenvalue optimization. Here we discuss three of them which are relevant to following chapters.

Formulation 1

Consider the optimization problem:

$$\begin{aligned} \max_{x} \quad & x^T A x \\ \text{s.t.} \quad & x^T x = 1 \end{aligned} \tag{2.5}$$

The Lagrangian is

$$L(x, \lambda) = x^T A x - \lambda(x^T x - 1) \tag{2.6}$$

The optimality condition $\nabla L(x, \lambda) = 0$ is

$$\begin{cases} Ax = \lambda x \\ x^T x = 1 \end{cases} \tag{2.7}$$

This is exactly the eigenvalue problem for matrix A. In other words, solving for the KKT system (or optimality conditions) is equivalent of solving for the eigenvalue problem. Multiplying the first equation of the system by x^T, we have $x^T A x = \lambda x^T x = \lambda$. Therefore, solving for the original optimization problem is equivalent of solving for the largest eigenvalue of matrix A.

Formulation 2

Consider the optimization problem:

$$\max_{x} \quad \frac{x^T A x}{x^T x} \tag{2.8}$$

It is easy to see that this formulation is equivalent to the previous optimization form. Therefore, this problem is also equivalent to the corresponding eigenvalue problem.

Formulation 3

Consider the optimization problem:

$$\begin{aligned} \max_{X} \quad & \mathrm{Tr}(X^T A X) \\ \text{s.t.} \quad & X^T X = I \end{aligned} \tag{2.9}$$

where the decision variable $X \in R^{n \times k}$ is a matrix.
The Lagrangian[14] is

$$L(X, \Lambda) = \mathrm{Tr}(X^T A X) - \mathrm{Tr}(\Lambda^T (X^T X - I)) \tag{2.10}$$

From matrix calculus, we know that

$$\frac{\partial \mathrm{Tr}(X^T A X)}{\partial X} = 2AX \tag{2.11}$$

The optimality condition $\nabla L(X, \Lambda) = 0$ is

$$\begin{cases} AX = X\Lambda \\ X^T X = I \end{cases} \tag{2.12}$$

This is exactly the eigenvalue problem for matrix A. In other words, solving for the KKT system (or optimality conditions) is equivalent of solving for the eigenvalue problem. Multiplying the first equation of the system by X^T, we have $X^T A X = X^T X \Lambda = \Lambda$ and $\mathrm{Tr}(X^T A X) = \mathrm{Tr}(\Lambda) = \Sigma_{i=1}^{k} \lambda_i$. Therefore, solving for the original optimization problem is equivalent of solving for the k largest eigenvalue of matrix A. It is easy to see that formulation 1 is a special case of this optimization form when $k = 1$.

2.4 Generalized Eigenvalue Optimization

Similarly, symmetric generalized eigenvalue problem is closely related to some specific optimization problems. [21] summarizes five forms of generalized eigenvalue optimization. Here we discuss three of them which are relevant to following chapters.

Formulation 1

Consider the optimization problem:

$$\begin{aligned} \max_{x} \quad & x^T A x \\ \text{s.t.} \quad & x^T B x = 1 \end{aligned} \tag{2.13}$$

The Lagrangian is

$$L(x, \lambda) = x^T Ax - \lambda(x^T Bx - 1) \tag{2.14}$$

The optimality condition $\nabla L(x, \lambda) = 0$ is

$$\begin{cases} Ax = \lambda Bx \\ x^T Bx = 1 \end{cases} \tag{2.15}$$

Note that the direction of x only depends on the first equation and the second equation is just for determining the scale of x. Therefore, this is exactly the generalized eigenvalue problem for matrix pair (A, B). In other words, solving for the KKT system (or optimality conditions) is equivalent of solving for the generalized eigenvalue problem. Multiplying the first equation of the system by x^T, we have $x^T Ax = \lambda x^T Bx = \lambda$. Therefore, solving for the original optimization problem is equivalent of solving for the largest generalized eigenvalue of (A, B).

Formulation 2

Consider the optimization problem:

$$\max_x \quad \frac{x^T Ax}{x^T Bx} \tag{2.16}$$

It is easy to see that this formulation is equivalent to the previous optimization form. Therefore, this problem is also equivalent to the corresponding generalized eigenvalue problem.

Formulation 3

Consider the optimization problem:

$$\begin{aligned} \max_X \quad & \mathrm{Tr}(X^T AX) \\ \text{s.t.} \quad & X^T BX = I \end{aligned} \tag{2.17}$$

where the decision variable $X \in R^{n \times k}$ is a matrix.
The Lagrangian[14] is

$$L(X, \Lambda) = \mathrm{Tr}(X^T AX) - \mathrm{Tr}(\Lambda^T (X^T BX - I)) \tag{2.18}$$

The optimality condition $\nabla L(X, \Lambda) = 0$ is

$$\begin{cases} AX = BX\Lambda \\ X^T BX = I \end{cases} \tag{2.19}$$

Though not obvious, the column span of X only depends on the first equation and the second equation is just for determining the scale of columns of X. Therefore this is exactly the generalized eigenvalue problem for matrix pair (A, B). In other words, solving for the KKT system (or optimality conditions) is equivalent of solving for the generalized eigenvalue problem. Multiplying the first equation of the system by X^T, we have $X^TAX = X^TBX\Lambda = \Lambda$ and $\mathrm{Tr}(X^TAX) = \mathrm{Tr}(\Lambda) = \Sigma_{i=1}^k \lambda_i$. Therefore, solving for the original optimization problem is equivalent of solving for the k largest generalized eigenvalue of (A, B). It is easy to see that formulation 1 is a special case of this optimization form when $k = 1$.

2.5 Examples

[21] introduces several examples of eigenvalue optimization and generalized eigenvalue optimization. We briefly introduce two of them. We also introduce one more example which is also relevant.

Kernel Supervised Principle Component Analysis

Kernel supervised PCA (SPCA) solves the following optimization:

$$\begin{aligned} \max_{\Theta} \quad & \mathrm{Tr}(\Theta^T(K_xHK_yHK_x)\Theta) \\ \text{s.t.} \quad & \Theta^TK_x\Theta = I \end{aligned} \tag{2.20}$$

where K_x and K_y are the kernel matrices over the training data and the labels, respectively, and $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the centering matrix. The goal is to find the kernel SPCA subspace denoted by Θ.

As discussed, the optimality condition is:

$$\begin{cases} (K_xHK_yHK_x)X = K_x\Theta\Lambda \\ \Theta^TK_x\Theta = I \end{cases} \tag{2.21}$$

which is a generalized eigenvalue problem.

Fisher Discriminant Analysis

Fisher Discriminant Analysis, or linear discriminant analysis (LDA), solves the following optimization:

$$\begin{aligned} \max_{X} \quad & \mathrm{Tr}(X^TS_bX) \\ \text{s.t.} \quad & X^TS_wX = I \end{aligned} \tag{2.22}$$

where S_b and S_w are the between-class covariance matrix and the within-class covariance matrix, respectively. The goal is to find the projection subspace denoted by X for dimension reduction. The mathematical definitions of S_b and S_w are:

$$S_b = \frac{1}{N}\Sigma_{i=1}^{N}(\mu_i - \mu)(\mu_i - \mu)^T \tag{2.23}$$

$$S_w = \frac{1}{N}\Sigma_{i=1}^{N}S_i \tag{2.24}$$

where μ_i and S_i are the mean vector and the class-specific covariance matrix of the i^{th} class of data points.

As discussed, the optimality condition is:

$$\begin{cases} S_bX = S_wX\Lambda \\ X^TS_wX = I \end{cases} \tag{2.25}$$

which is a generalized eigenvalue problem.

Adaptive Beamforming

Adaptive beamforming solves the following optimization problem:

$$\max_w \quad \frac{w^HR_sw}{w^HR_{i+n}w} \tag{2.26}$$

where w is the vector of beamformer weights, and R_s and R_{i+n} are the signal and interference-plus-noise covariance matrices respectively. The goal is to find the beamformer weight w such that the signal-to-interference-plus-noise ratio(SINR) is maximized.

As discussed, the optimality condition is:

$$\begin{cases} R_sw = \lambda R_{i+n}w \\ w^HR_{i+n}w = 1 \end{cases} \tag{2.27}$$

which is a generalized eigenvalue problem.

Chapter 3

Robust GEP: Rank 1

3.1 Formulations

In this section, we study different formulations of robust generalized eigenvalue problem for positive semi-definite matrices. The key difference is how we model data uncertainty in positive semi-definite matrices. For a general matrix A, if we believe that the data is uncertain, we can assume that the true matrix lies in the following set:

$$S_A = \{A + \Delta A \mid \|\Delta A\| \leq \epsilon\}$$

However, matrices in S_A are not necessarily positive semi-definite even when A is positive semi-definite. In the cases where the true matrix is known to be positive semi-definite and $A = \tilde{A}\tilde{A}^T$, we can assume that the true matrix lies in the following set:

$$\tilde{S}_A = \{(\tilde{A} + \Delta\tilde{A})(\tilde{A} + \Delta\tilde{A})^T \mid \|\Delta\tilde{A}\| \leq \epsilon\} \tag{3.1}$$

Matrices in $\tilde{S}_A$ are always positive semi-definite , so this set well describes positive semi-definite matrices that are close to A. Note that $\tilde{S}_A$ is not dependent on the decomposition $A = \tilde{A}\tilde{A}^T$.

Proposition 1. *The set $\tilde{S}_A$ defined in (3.1) is independent of the decomposition $A = \tilde{A}\tilde{A}^T$*

Proof. Suppose $A = \tilde{A}_1\tilde{A}_1^T = \tilde{A}_2\tilde{A}_2^T$, then we have

$$(\tilde{A}_1^T\tilde{A}_1)\tilde{A}_1^T = (\tilde{A}_1^T\tilde{A}_2)\tilde{A}_2^T$$

Since $\tilde{A}_1$ is assumed to be full rank, $\tilde{A}_1^T\tilde{A}_1$ is invertible and

$$\tilde{A}_1^T = (\tilde{A}_1^T\tilde{A}_1)^{-1}(\tilde{A}_1^T\tilde{A}_2)\tilde{A}_2^T$$

Let $Q = (\tilde{A}_1^T \tilde{A}_1)^{-1}(\tilde{A}_1^T \tilde{A}_2)$, then

$$QQ^T = (\tilde{A}_1^T \tilde{A}_1)^{-1}(\tilde{A}_1^T \tilde{A}_2)(\tilde{A}_1^T \tilde{A}_2)^T(\tilde{A}_1^T \tilde{A}_1)^{-T} = I$$

so Q is orthogonal.
Now for any element u in $\tilde{S}_{A_1} = \{(\tilde{A}_1 + \Delta\tilde{A}_1)(\tilde{A}_1 + \Delta\tilde{A}_1)^T \mid \|\Delta\tilde{A}_1\| \leq \epsilon\}$ defined based on $\tilde{A}_1$, it can be written as

$$\begin{aligned} u &= (\tilde{A}_1 + \Delta\tilde{A}_1)(\tilde{A}_1 + \Delta\tilde{A}_1)^T \\ &= (\tilde{A}_2 Q + \Delta\tilde{A}_1)(\tilde{A}_2 Q + \Delta\tilde{A}_1)^T \\ &= (\tilde{A}_2 + \Delta\tilde{A}_1 Q^{-1})QQ^T(\tilde{A}_2 + \Delta\tilde{A}_1 Q^{-1})^T \\ &= (\tilde{A}_2 + \Delta\tilde{A}_1 Q^{-1})(\tilde{A}_2 + \Delta\tilde{A}_1 Q^{-1})^T \end{aligned}$$

Since $\|\Delta\tilde{A}_1 Q^{-1}\| = \|\Delta\tilde{A}_1\| \leq \epsilon$, $u \in \tilde{S}_{A_2} = \{(\tilde{A}_2 + \Delta\tilde{A}_2)(\tilde{A}_2 + \Delta\tilde{A}_2)^T \mid \|\Delta\tilde{A}_2\| \leq \epsilon\}$ defined based on $\tilde{A}_2$. So $\tilde{S}_{A_1} \subseteq \tilde{S}_{A_2}$. By symmetry, $\tilde{S}_{A_1} = \tilde{S}_{A_2}$ □

From section 2, we know that the leading generalized eigenvector can be defined as a solution to the following optimization problem:

$$\max_x \quad \frac{x^T A x}{x^T B x}$$

Depending on how we model data uncertainty in positive definite matrices, we have three different robust formulations.

Formulation 1

When we use the first type of uncertainty set for both A and B, we can formulate the robust optimization problem as:

$$\max_x \min_{\|\Delta A\| \leq \epsilon_A, \|\Delta B\| \leq \epsilon_B} \quad \frac{x^T(A + \Delta A)x}{x^T(B + \Delta B)x} \tag{3.2}$$

Since A and B are positive definite, when ϵ_A and ϵ_B are small enough, both numerator and denominator are positive, so the problem is equivalent to the following:

$$\max_x \quad \frac{\min_{\|\Delta A\| \leq \epsilon_A} x^T(A + \Delta A)x}{\max_{\|\Delta B\| \leq \epsilon_B} x^T(B + \Delta B)x} \tag{3.3}$$

Then we need to solve these two subproblems

Lemma 1. *The optimal value for* $\min_{\|\Delta A\| \le \epsilon_A} x^T(A+\Delta A)x$ *is* $x^T(A-\epsilon_A I)x$ *and the optimal value for* $\max_{\|\Delta B\| \le \epsilon_B} x^T(B+\Delta B)x$ *is* $x^T(B+\epsilon_B I)x$

Proof. We can always choose $\Delta A = -\epsilon_A x x^T$ and $\Delta B = \epsilon_B x x^T$, where $\|\Delta A\| = \epsilon_A$ and $\|\Delta B\| = \epsilon_B$, and then

$$x^T(A+\Delta A)x = x^T(A-\epsilon_A I)x$$

$$x^T(B+\Delta B)x = x^T(B+\epsilon_B I)x$$

On the other hand,

$$x^T \Delta A x \ge -\|x^T\|_2 \|\Delta A\|_2 \|x\|_2 \ge -\epsilon_A x^T x$$

$$x^T \Delta B x \le \|x^T\|_2 \|\Delta B\|_2 \|x\|_2 \le \epsilon_B x^T x$$

This completes the proof. □

Then the problem becomes

$$\max_x \quad \frac{x^T(A-\epsilon_A I)x}{x^T(B+\epsilon_B I)x} \tag{3.4}$$

Now we can solve this optimization problem directly. This is nothing but a generalized eigenvalue problem of $(A-\epsilon_A I, B+\epsilon_B I)$

Formulation 2

When we use the first type of uncertainty set for B and the second type of uncertainty set for A, we can formulate the robust optimization problem as:

$$\max_x \min_{\|\Delta\tilde{A}\| \le \epsilon_A, \|\Delta B\| \le \epsilon_B} \quad \frac{\|(\tilde{A}+\Delta\tilde{A})x\|_2^2}{x^T(B+\Delta B)x} \tag{3.5}$$

Again, since both numerator and denominator are positive, the problem is equivalent to the following:

$$\max_x \quad \frac{\min_{\|\Delta\tilde{A}\| \le \epsilon_A} \|(\tilde{A}+\Delta\tilde{A})x\|_2^2}{\max_{\|\Delta B\| \le \epsilon_B} x^T(B+\Delta B)x} \tag{3.6}$$

Then we need to solve these two subproblems

Lemma 2. *The optimal value for* $\min_{\|\Delta\tilde{A}\| \le \epsilon_A} \|(\tilde{A}+\Delta\tilde{A})x\|_2$ *is* $\|\tilde{A}x\|_2 - \epsilon_A\|x\|_2$

Proof. When $\|\tilde{A}x\|_2 \leq \epsilon_A \|x\|_2$, we can choose

$$\Delta\tilde{A} = -\frac{\tilde{A}xx^T}{\|x\|_2^2}$$

where

$$\|\Delta\tilde{A}\| = \frac{\|\tilde{A}xx^T\|_2}{\|x\|_2^2} \leq \frac{\|\tilde{A}x\|_2\|x^T\|_2}{\|x\|_2^2} = \frac{\|\tilde{A}x\|_2}{\|x\|_2} \leq \epsilon_A$$

Then

$$\|(\tilde{A} + \Delta\tilde{A})x\| = \|(\tilde{A} - \frac{\tilde{A}xx^T}{\|x\|_2^2})x\| = \|\tilde{A}x - \frac{\tilde{A}xx^Tx}{\|x\|_2^2}\| = \|\tilde{A}x - \tilde{A}x\| = 0$$

When $\|\tilde{A}x\|_2 > \epsilon_A \|x\|_2$, we can choose

$$\Delta\tilde{A} = -\frac{\epsilon_A\tilde{A}xx^T}{\|\tilde{A}x\|_2\|x\|_2}$$

where

$$\|\Delta\tilde{A}\| = \epsilon_A\frac{\|\tilde{A}xx^T\|_2}{\|\tilde{A}x\|_2\|x\|_2} \leq \epsilon_A\frac{\|\tilde{A}x\|_2\|x^T\|_2}{\|\tilde{A}x\|_2\|x\|_2} = \epsilon_A$$

Then

$$\begin{aligned}
\|(\tilde{A} + \Delta\tilde{A})x\| &= \|(\tilde{A} - \epsilon_A\frac{\tilde{A}xx^T}{\|\tilde{A}x\|_2\|x\|_2})x\| \\
&= \|\tilde{A}x - \epsilon_A\frac{\tilde{A}xx^Tx}{\|\tilde{A}x\|_2\|x\|_2}\| \\
&= \|\tilde{A}x - \epsilon_A\frac{\|x\|_2}{\|\tilde{A}x\|_2}\tilde{A}x\| \\
&= (1 - \epsilon_A\frac{\|x\|_2}{\|\tilde{A}x\|_2})\|\tilde{A}x\|_2 \\
&= \|\tilde{A}x\|_2 - \epsilon_A\|x\|_2
\end{aligned}$$

On the other hand,

$$\|(\tilde{A} + \Delta\tilde{A})x\| \geq 0$$

and

$$\|(\tilde{A} + \Delta\tilde{A})x\| \geq \|\tilde{A}x\| - \|\Delta\tilde{A}x\| \geq \|\tilde{A}x\| - \|\Delta\tilde{A}\|\|x\| \geq \|\tilde{A}x\| - \epsilon_A\|x\|$$

Therefore

$$\|(\tilde{A} + \Delta\tilde{A})x\| \geq \max(0, \|\tilde{A}x\| - \epsilon_A\|x\|)$$

We already showed that the equality can be achieved. This completes the proof. □

Then the problem becomes

$$\max_x \quad \frac{(\|\tilde{A}x\|_2 - \epsilon_A\|x\|_2)^2}{x^T(B+\epsilon_B I)x} \tag{3.7}$$

Let $y = Cx$, where $B + \epsilon_B = C^T C$, then the problem becomes

$$\max_y \quad \frac{(\|\tilde{A}C^{-1}y\|_2 - \epsilon_A\|C^{-1}y\|_2)^2}{y^T y} \tag{3.8}$$

This can be reformulated as a constrained optimization problem:

$$\begin{aligned} \max_y \quad & \|Fy\|_2 - \|Gy\|_2 \\ \text{s.t.} \quad & \|y\|_2 = 1 \end{aligned} \tag{3.9}$$

This is a non-convex optimization problem. We will solve for this problem in the next section.

Formulation 3

When we use the second type of uncertainty set for both A and B, we can formulate the robust optimization problem as:

$$\max_x \min_{\|\Delta\tilde{A}\|\le\epsilon_A, \|\Delta\tilde{B}\|\le\epsilon_B} \quad \frac{\|(\tilde{A}+\Delta\tilde{A})x\|_2}{\|(\tilde{B}+\Delta\tilde{B})x\|_2} \tag{3.10}$$

Again, since both numerator and denominator are positive, the problem is equivalent to the following:

$$\max_x \quad \frac{\min_{\|\Delta\tilde{A}\|\le\epsilon_A} \|(\tilde{A}+\Delta\tilde{A})x\|_2^2}{\max_{\|\Delta\tilde{B}\|\le\epsilon_B} \|(\tilde{B}+\Delta\tilde{B})x\|_2^2} \tag{3.11}$$

Then we need to solve these two subproblems

Lemma 3. *The optimal value for* $\max_{\|\Delta\tilde{B}\|\le\epsilon_B} \|(\tilde{B}+\Delta\tilde{B})x\|_2$ *is* $\|\tilde{B}x\|_2 + \epsilon_B\|x\|_2$

Then the problem becomes

$$\max_x \quad \frac{\|\tilde{A}x\|_2 - \epsilon_A\|x\|_2}{\|\tilde{B}x\|_2 + \epsilon_B\|x\|_2} \tag{3.12}$$

This can be reformulated as a constrained optimization problem:

$$\begin{aligned} \max_x \quad & \|\tilde{A}x\|_2 - \epsilon_A\|x\|_2 \\ \text{s.t.} \quad & \|\tilde{B}x\|_2 + \epsilon_B\|x\|_2 = 1 \end{aligned} \tag{3.13}$$

This is a non-convex optimization problem. We will solve for this problem in the next section.

Comments

When the matrix A is not full rank, the difference between formulation 1 (3.2) and the other two formulations is substantially different since $A + \Delta A$ is no longer a positive semi-definite matrix in formulation 1. However, the difference between formulation 2 (3.5) and formulation 3 (3.10) is subtle, since $B + \Delta B$ is still a positive semi-definite matrix in formulation 2. However, formulation 1 leads to a closed-form solution while formulation 2 and 3 require more computational resources.

3.2 Optimality Conditions

In this section, we derive the optimality conditions for the two non-convex optimization problems mentioned in the previous section.

Formulation 2

Consider the optimization problem:

$$\begin{aligned} \max_{x} \quad & \|Fx\|_2 - \|Gx\|_2 \\ \text{s.t.} \quad & \|x\|_2 = 1 \end{aligned} \tag{3.14}$$

The Lagrangian is

$$L(x, \lambda) = \|Fx\|_2 - \|Gx\|_2 - \lambda(\|x\|_2 - 1) \tag{3.15}$$

The optimality condition $\nabla L(x, \lambda) = 0$ is

$$\begin{cases} \dfrac{F^T F x}{\|Fx\|_2} - \dfrac{G^T G x}{\|Gx\|_2} = \lambda \dfrac{x}{\|x\|_2} \\ \|x\|_2 = 1 \end{cases} \tag{3.16}$$

We can solve for λ in terms of x by multiplying x by the left on both sides of the first equation:

$$\|Fx\|_2 - \|Gx\|_2 = \lambda \|x\|_2 = \lambda \tag{3.17}$$

Then the optimality condition becomes

$$\begin{cases} \dfrac{F^T F x}{\|Fx\|_2} - \dfrac{G^T G x}{\|Gx\|_2} = (\|Fx\|_2 - \|Gx\|_2) \dfrac{x}{\|x\|_2} \\ \|x\|_2 = 1 \end{cases} \tag{3.18}$$

It is difficult for us to solve for this highly non-linear system of equations directly. Instead, we introduce auxiliary variables α, β, μ and aim at solving for a relaxed system of equations. Specifically, we are looking for (α, β, μ, x) such that

$$\begin{cases} \dfrac{F^TFx}{\alpha} - \dfrac{G^TGx}{\beta} = \mu x \\ \alpha - \beta = \mu \\ ||x||_2 = 1 \\ \alpha, \beta > 0 \end{cases} \tag{3.19}$$

This is a non linear system of equations that is easier to solve than the original one and it is closely related to the optimality condition. First, for any solution (λ, x) to the optimality condition, it is also a solution to this relaxed system of equations. This is because we can simply set $\alpha = \|Fx\|_2$, $\beta = \|Gx\|_2$, $\mu = \lambda$, and x unchanged. Note that the other direction is not always true, which means that not all solutions (α, β, μ, x) to the relaxed system of equations could lead to a solution (λ, x) to the original optimality condition. However, the specific solution that maximizes μ, which we are mostly interested in, would guarantee a solution to the optimality condition.

Theorem 1. *Let P be the solution path of the system of equations (3.19), then $\lambda = \max\{\mu | (\alpha, \beta, \mu, x) \in P\}$ and the corresponding unit vector x is the solution to the optimality condition (3.16).*

Proof. Let $(\alpha + \Delta\alpha, \beta + \Delta\beta, \mu + \Delta\mu, x + \Delta x)$ be a perturbed solution to (α, β, μ, x). Then we have

$$\frac{F^TFx}{\alpha} - \frac{G^TGx}{\beta} = \mu x \tag{3.20}$$

$$\frac{F^TF(x+\Delta x)}{\alpha+\Delta\alpha} - \frac{G^TG(x+\Delta x)}{\beta+\Delta\beta} = (\mu+\Delta\mu)(x+\Delta x) \tag{3.21}$$

Taking the difference, we have

$$(\frac{F^TF\Delta x}{\alpha} - \frac{G^TG\Delta x}{\beta}) - (\frac{F^TFx}{\alpha^2}\Delta\alpha - \frac{G^TGx}{\beta^2}\Delta\beta) = \mu\Delta x + \Delta\mu x \tag{3.22}$$

Multiplying x^T by the left, we have

$$-(\frac{\|Fx\|_2^2}{\alpha^2}\Delta\alpha - \frac{\|Gx\|_2^2}{\beta^2}\Delta\beta) = \Delta\mu \tag{3.23}$$

Note that $\Delta\beta = \Delta\alpha - \Delta\mu$, therefore

$$(1 + \frac{\|Gx\|_2^2}{\beta^2})\Delta\mu = -\Delta\alpha(\frac{\|Fx\|_2^2}{\alpha^2} - \frac{\|Gx\|_2^2}{\beta^2}) \tag{3.24}$$

For the specific solution $\lambda = \max\{\mu|(\alpha, \beta, \mu, x) \in P\}$, $\frac{\Delta\mu}{\Delta\alpha}$ should be zero, otherwise we could perturb α in the direction such that μ increases. Therefore

$$\frac{\|Fx\|_2}{\alpha} = \frac{\|Gx\|_2}{\beta} \tag{3.25}$$

From the original equations, we have

$$\frac{\|Fx\|_2^2}{\alpha} - \frac{\|Gx\|_2^2}{\beta} = \mu = \alpha - \beta \tag{3.26}$$

Therefore

$$\begin{cases} \alpha = \|Fx\|_2 \\ \beta = \|Gx\|_2 \\ \mu = \|Fx\|_2 - \|Gx\|_2 \end{cases} \tag{3.27}$$

In other words, $\lambda = \max\{\mu|(\alpha, \beta, \mu, x) \in P\}$ and the corresponding unit vector x is the solution to the optimality condition. □

The theorem basically says that we can solve for the relaxed system of equations to obtain the maximum λ solution to the optimality condition. In the next section, we will discuss how we can solve the relaxed system of equations in detail.

Formulation 3

Consider the optimization problem:

$$\begin{aligned} \max_x \quad & \|\tilde{A}x\|_2 - \epsilon_A\|x\|_2 \\ \text{s.t.} \quad & \|\tilde{B}x\|_2 + \epsilon_B\|x\|_2 = 1 \end{aligned} \tag{3.28}$$

The Lagrangian is

$$L(x, \lambda) = \|\tilde{A}x\|_2 - \epsilon_A\|x\|_2 - \lambda(\|\tilde{B}x\|_2 + \epsilon_B\|x\|_2 - 1) \tag{3.29}$$

The optimality condition $\nabla L(x, \lambda) = 0$ is

$$\begin{cases} \frac{\tilde{A}^T\tilde{A}x}{\|\tilde{A}x\|_2} - \epsilon_A\frac{x}{\|x\|_2} = \lambda(\frac{\tilde{B}^T\tilde{B}x}{\|\tilde{B}x\|_2} + \epsilon_B\frac{x}{\|x\|_2}) \\ \|\tilde{B}x\|_2 + \epsilon_B\|x\|_2 = 1 \end{cases} \tag{3.30}$$

We can solve for λ in terms of x by multiplying x by the left on both sides of the first equation:

$$\|\tilde{A}x\|_2 - \epsilon_A\|x\|_2 = \lambda(\|\tilde{B}x\|_2 + \epsilon_B\|x\|_2) = \lambda \tag{3.31}$$

Then the optimality condition becomes

$$\begin{cases} \dfrac{\tilde{A}^T\tilde{A}x}{\|\tilde{A}x\|_2} - \epsilon_A\dfrac{x}{\|x\|_2} = (\|\tilde{A}x\|_2 - \epsilon_A\|x\|_2)(\dfrac{\tilde{B}^T\tilde{B}x}{\|\tilde{B}x\|_2} + \epsilon_B\dfrac{x}{\|x\|_2}) \\ \|\tilde{B}x\|_2 + \epsilon_B\|x\|_2 = 1 \end{cases} \tag{3.32}$$

It is difficult for us to solve for this highly non-linear system of equations directly. Instead, we introduce auxiliary variables $\alpha, \beta, \gamma, \mu$ and aim at solving for a relaxed system of equations. Specifically, we are looking for $(\alpha, \beta, \gamma, \mu, x)$ such that

$$\begin{cases} \dfrac{\tilde{A}^T\tilde{A}x}{\alpha} - \epsilon_A x = \mu(\dfrac{\tilde{B}^T\tilde{B}x}{\beta} + \epsilon_B x) \\ (\alpha - \epsilon_A)\gamma = \mu \\ (\beta + \epsilon_B)\gamma = 1 \\ \|x\|_2 = 1 \\ \alpha, \beta, \gamma > 0 \end{cases} \tag{3.33}$$

Eliminating the scale factor λ, we have the simplified system of equations

$$\begin{cases} \dfrac{\tilde{A}^T\tilde{A}x}{\alpha} - \epsilon_A x = \mu(\dfrac{\tilde{B}^T\tilde{B}x}{\beta} + \epsilon_B x) \\ \alpha - \epsilon_A = \mu(\beta + \epsilon_B) \\ \|x\|_2 = 1 \\ \alpha, \beta > 0 \end{cases} \tag{3.34}$$

Again, this is a non linear system of equations that is easier to solve than the original one and it is closely related to the optimality condition. First, for any solution $(\lambda, \tilde{x})$ to the optimality condition, it is also a solution to this relaxed system of equations. This is because we can simply set $\alpha = \frac{\|\tilde{A}\tilde{x}\|_2}{\|\tilde{x}\|_2}$, $\beta = \frac{\|\tilde{B}\tilde{x}\|_2}{\|\tilde{x}\|_2}$, $\mu = \lambda$, and $x = \frac{\tilde{x}}{\|\tilde{x}\|_2}$. Note that the other direction is not always true, which means that not all solutions (α, β, μ, x) to the relaxed system of equations could lead to a solution (λ, x) to the original optimality condition. However, the specific solution that maximizes μ, which we are mostly interested in, would guarantee a solution to the optimality condition.

Theorem 2. *Let P be the solution path of the above system of equations, then $\lambda = \max\{\mu|(\alpha, \beta, \mu, x) \in P\}$ and the corresponding unit vector x is the solution to the optimality condition.*

Proof. Let $(\alpha+\Delta\alpha, \beta+\Delta\beta, \mu+\Delta\mu, x+\Delta x)$ be a perturbed solution to (α, β, μ, x). Then we have

$$\frac{\tilde{A}^T\tilde{A}x}{\alpha} - \epsilon_A x = \mu(\frac{\tilde{B}^T\tilde{B}x}{\beta} + \epsilon_B x) \tag{3.35}$$

$$\frac{\tilde{A}^T\tilde{A}(x+\Delta x)}{\alpha+\Delta\alpha} - \epsilon_A(x+\Delta x) = (\mu+\Delta\mu)(\frac{\tilde{B}^T\tilde{B}(x+\Delta x)}{\beta+\Delta\beta} + \epsilon_B(x+\Delta x)) \tag{3.36}$$

Taking the difference, we have

$$(\frac{\tilde{A}^T\tilde{A}}{\alpha} - \mu\frac{\tilde{B}^T\tilde{B}}{\beta})\Delta x - (\frac{\tilde{A}^T\tilde{A}x}{\alpha^2}\Delta\alpha - \mu\frac{\tilde{B}^T\tilde{B}x}{\beta^2}\Delta\beta) = (\epsilon_A + \mu\epsilon_B)\Delta x + \Delta\mu(\frac{\tilde{B}^T\tilde{B}x}{\beta} + \epsilon_B x) \tag{3.37}$$

Multiplying x^T by the left, we have

$$-(\frac{\|\tilde{A}x\|_2^2}{\alpha^2}\Delta\alpha - \mu\frac{\|\tilde{B}x\|_2^2}{\beta^2}\Delta\beta) = \Delta\mu(\frac{\|\tilde{B}x\|_2^2}{\beta} + \epsilon_B) \tag{3.38}$$

Note that $\Delta\alpha = (\beta+\epsilon_B)\Delta\mu + \mu\Delta\beta$, therefore

$$-\mu(\frac{\|\tilde{A}x\|_2^2}{\alpha^2} - \frac{\|\tilde{B}x\|_2^2}{\beta^2})\Delta\beta = \Delta\mu(\frac{\|\tilde{B}x\|_2^2}{\beta} + \epsilon_B + (\beta+\epsilon_B)\frac{\|\tilde{A}x\|_2^2}{\alpha^2}) \tag{3.39}$$

For the specific solution $\lambda = \max\{\mu|(\alpha, \beta, \mu, x) \in P\}$, $\frac{\Delta\mu}{\Delta\beta}$ should be zero, otherwise we could perturb β in the direction such that μ increases. Therefore

$$\frac{\|\tilde{A}x\|_2}{\alpha} = \frac{\|\tilde{B}x\|_2}{\beta} \tag{3.40}$$

Now we have

$$\begin{cases} \dfrac{\|\tilde{A}x\|_2^2}{\alpha} - \epsilon_A = \mu(\dfrac{\|\tilde{B}x\|_2^2}{\beta} + \epsilon_B) \\ \alpha - \epsilon_A = \mu(\beta + \epsilon_B) \\ \dfrac{\|\tilde{A}x\|_2}{\alpha} = \dfrac{\|\tilde{B}x\|_2}{\beta} \end{cases} \tag{3.41}$$

Therefore

$$\begin{cases} \alpha = \|\tilde{A}x\|_2 \\ \beta = \|\tilde{B}x\|_2 \\ \mu = \dfrac{\|\tilde{A}x\|_2 - \epsilon_A}{\|\tilde{B}x\|_2 + \epsilon_B} \end{cases} \tag{3.42}$$

In other words, $\lambda = \max\{\mu|(\alpha, \beta, \mu, x) \in P\}$ and the corresponding unit vector x is the solution to the optimality condition. □

The theorem basically says that we can solve for the relaxed system of equations to obtain the maximum λ solution to the optimality condition. In the next section, we will discuss how we can solve the relaxed system of equations in detail.

3.3 Algorithms

In this section, we introduced two algorithms to solve the above non-linear system of equations.

Consider the nonlinear system of equations(3.19), after eliminating α, we have

$$\begin{cases} \dfrac{\tilde{A}^T\tilde{A}x}{\beta+\mu} - \dfrac{\tilde{B}^T\tilde{B}x}{\beta} = \mu x \\ ||x||_2 = 1 \\ \beta > 0 \end{cases} \tag{3.43}$$

There could be more than one set of solutions (β, μ, x) and our goal is to find the solution with maximal μ. Given μ, this is in fact a quadratic eigenvalue problem of (β, x) with an additional condition $\beta > 0$:

$$\begin{cases} (\beta^2\mu I + \beta(\mu^2 I - \tilde{A}^T\tilde{A} + \tilde{B}^T\tilde{B}) + \mu\tilde{B}^T\tilde{B})x = 0 \\ \|x\|_2 = 1 \\ \beta > 0 \end{cases} \tag{3.44}$$

Similarly, we can consider the nonlinear system of equations(3.34), after eliminating α, we have

$$\begin{cases} \dfrac{\tilde{A}^T\tilde{A}x}{\epsilon_A + \mu(\beta+\epsilon_B)} - r_A x = \mu(\dfrac{\tilde{B}^T\tilde{B}x}{\beta} + \epsilon_B x) \\ \|x\|_2 = 1 \\ \beta > 0 \end{cases} \tag{3.45}$$

There could be more than one set of solutions (β, μ, x) and our goal is to find the solution with maximal μ. Given μ, this is in fact a quadratic eigenvalue problem of (β, x) with an additional condition $\beta > 0$:

$$\begin{cases} (\beta^2 c\mu I + \beta(c^2 I - \tilde{A}^T\tilde{A} + \mu^2\tilde{B}^T\tilde{B}) + c\mu\tilde{B}^T\tilde{B})x = 0 \\ \|x\|_2 = 1 \\ \beta > 0 \end{cases} \tag{3.46}$$

where $c = r_A + \mu r_B$
(3.44) and (3.46) suggest the following root finding algorithm for solving for the optimal (β^*, μ^*, x^*):

Step 1: Given μ, test whether (3.44) has solution or not

Step 2: If yes, set $\mu_{new} > \mu$, otherwise, set $\mu_{new} < \mu$

Step 3: Iterate until the stop criterion is satisfied

In step 1, we solve for the quadratic eigenvalue problem (QEP) given μ. There would be $2n$ solution pairs $(\beta(\mu), x(\mu))$ and we would be interested in whether there are any solution pair such that $\beta(\mu) > 0$. If this is true, the equation (3.44)(or(3.46)) has a feasible solution tuple (β, μ, x), and this means that the maximal μ^* among all feasible solutions would be greater than or equal to μ. Therefore in step 2, we test another μ_{new} that is larger than μ. On the other hand, if none of the solution satisfies $\beta(\mu) > 0$, this means that that the maximal μ^* among all feasible solutions would be less than μ. This is because μ is a continuous function of β. Therefore in step 2, we test another μ_{new} that is less than μ. This way, after every iteration, μ_{new} gets closer and closer to the optimal value μ^* and we can use some stopping criteria to terminate the iteration.

The only question left is how we update μ_{new}. There are many ways to update μ_{new}. Now we introduce two methods to update μ_{new}.

Bisection

It is obvious that $\mu^* \geq 0$. When we have an upper bound of μ^*, which is generally true in practice, we can use bisection method to update μ_{new}. Algorithm 1 is the pseudocode of bisection method for solving (3.44). The QEP function in the pseudocode solves the following quadratic eigenvalue problem and returns all real eigenvalues β_k and corresponding eigenvectors x_k, in the order of descending β_k.

$$(\beta^2 \mu I + \beta(\mu^2 I - \tilde{A}^T \tilde{A} + \tilde{B}^T \tilde{B}) + \mu \tilde{B}^T \tilde{B})x = 0 \tag{3.47}$$

Algorithm 1 Bisection

Input: matrix A, B, param $\epsilon_A, \epsilon_B, ub$
$lb, ub, i = 0, ub, 0$
while $i <$ MAXITER:
  $\mu = (lb + ub)/2$
  $\beta_k, x_k = \text{QEP}(A, B, \epsilon_A, \epsilon_B, \mu)$
  if $\beta_1 > 0$:
    $lb, x^* = \mu, x_1$
  else:
    $ub = \mu$
  $i = i + 1$
$\mu^* = (lb + ub)/2$
Output: optimal value μ^*, solution x^*

Boyd-Balakrishnan Method

Inspired by the quadratic convergent algorithm for computing the H_∞ norm of a matrix [13], we proposed another update scheme for μ_{new}. Algorithm 2 is the pseudocode of the Boyd-Balakrishnan method for solving (3.44). The QEP function is the same as that in Bisection and the QEP2 function solves the following quadratic eigenvalue problem and returns all real eigenvalues μ_k and corresponding eigenvectors x_k, in the order of descending μ_k.

$$(\mu^2\beta I + \mu(\beta^2 I + \tilde{B}^T\tilde{B}) + \beta(\tilde{B}^T\tilde{B} - \tilde{A}^T\tilde{A}))x = 0 \tag{3.48}$$

Experiments show that the Boyd-Balakrishnan method converges quadratically, which requires much less iterations compared to bisection. However, within each iteration, the Boyd-Balakrishnan method requires to solve multiple quadratic eigenvalue problems, while bisection method only solves one QEP. Therefore in practice, the Boyd-Balakrishnan method is not necessarily faster than bisection.

Algorithm 2 Boyd-Balakrishnan

Input: matrix A, B, param $\epsilon_A, \epsilon_B, \mu_0$
$\mu, i = \mu_0, 0$
while $i <$ MAXITER:
  $\beta_k, x_k = \text{QEP}(A, B, \epsilon_A, \epsilon_B, \mu)$
  for k in valid range:
    $\tilde{\beta}_k = (\beta_k + \beta_{k+1})/2$
    if $\tilde{\beta}_k > 0$:
      $\mu_k, x_k = \text{QEP2}(A, B, \epsilon_A, \epsilon_B, \tilde{\beta}_k)$
      $\mu = \max(\mu, \mu_1)$
  $i = i + 1$
$\beta_k, x_k = \text{QEP}(A, B, \epsilon_A, \epsilon_B, \mu)$
Output: optimal value μ, solution x_1

Chapter 4

Robust GEP: General rank

4.1 Formulation

Recall that there are two ways to model data uncertainty:

$$S_A = \{A + \Delta A \mid \|\Delta A\| \leq \epsilon\} \tag{4.1}$$

$$\tilde{S}_A = \{(\tilde{A} + \Delta\tilde{A})(\tilde{A} + \Delta\tilde{A})^T \mid \|\Delta\tilde{A}\| \leq \epsilon\} \tag{4.2}$$

From section 2, we know that the leading generalized eigenvectors can be defined as a solution to the following optimization problem:

$$\begin{aligned} \max_X \quad & \mathrm{Tr}(X^T A X) \\ \text{s.t.} \quad & X^T B X = I \end{aligned} \tag{4.3}$$

Depending on how we model data uncertainty in the positive semi-definite matrix A, we have two different robust formulations.

Formulation 1

When we use the first type of uncertainty set (4.1) for A, the robust optimization problem is:

$$\begin{aligned} \max_X \min_{\|\Delta A\| \leq \epsilon_A} \quad & \mathrm{Tr}(X^T (A + \Delta A) X) \\ \text{s.t.} \quad & X^T B X = I \end{aligned} \tag{4.4}$$

Then we need to solve the inner subproblem.

Lemma 4. *The optimal value for* $\min_{\|\Delta A\| \leq \epsilon_A} Tr(X^T(A + \Delta A)X)$ *is* $Tr(X^T A X) - \epsilon_A \|X^T X\|_F$

Proof. We can always choose $\Delta A = -\epsilon_A \frac{XX^T}{\|XX^T\|_F}$, where

$$\|\Delta A\| = \epsilon_A \frac{\|XX^T\|_F}{\|XX^T\|_F} = \epsilon_A$$

and then

$$\begin{aligned}
\mathrm{Tr}(X^T(A+\Delta A)X) &= \mathrm{Tr}(X^T(A - \epsilon_A \frac{XX^T}{\|XX^T\|_F})X) \\
&= \mathrm{Tr}(X^TAX - \epsilon_A \frac{X^TXX^TX}{\|XX^T\|_F}) \\
&= \mathrm{Tr}(X^TAX) - \epsilon_A \frac{\mathrm{Tr}(X^TXX^TX)}{\|XX^T\|_F} \\
&= \mathrm{Tr}(X^TAX) - \epsilon_A \frac{\|XX^T\|_F^2}{\|XX^T\|_F} \\
&= \mathrm{Tr}(X^TAX) - \epsilon_A \|XX^T\|_F \\
&= \mathrm{Tr}(X^TAX) - \epsilon_A \|XX^T\|_F
\end{aligned}$$

On the other hand,

$$\mathrm{Tr}(X^T\Delta AX) = \mathrm{Tr}(XX^T\Delta A) \geq -\|XX^T\|_F\|\Delta A\|_F \geq -\epsilon_A||XX^T||_F$$

This completes the proof. □

Then the problem becomes

$$\begin{aligned}
\max_X \quad & \mathrm{Tr}(X^TAX) - \epsilon_A\|X^TX\|_F \\
\text{s.t.} \quad & X^TBX = I
\end{aligned} \tag{4.5}$$

This is a non-convex optimization problem. We will analyze this problem in the next section.

Formulation 2

When we use the first type of uncertainty set for B and the second type of uncertainty set for A, we can formulate the robust optimization problem as:

$$\begin{aligned}
\max_X \min_{\|\Delta\tilde{A}\| \leq \epsilon_A} \quad & \mathrm{Tr}(X^T(\tilde{A}+\Delta\tilde{A})^T(\tilde{A}+\Delta\tilde{A})X) \\
\text{s.t.} \quad & X^TBX = I
\end{aligned} \tag{4.6}$$

Then we need to solve the inner sub-problem.

Lemma 5. *The optimal value for* $\min_{\|\Delta\tilde{A}\|\leq\epsilon_A} \quad Tr(X^T(\tilde{A}+\Delta\tilde{A})^T(\tilde{A}+\Delta\tilde{A})X)$ *is*

$$\begin{cases} \sigma^2\|\tilde{A}X(\sigma I + X^TX)^{-1}\|_F^2, & \|\tilde{A}X(X^TX)^{-1}X^T\|_F > \epsilon_A \\ 0, & \|\tilde{A}X(X^TX)^{-1}X^T\|_F \leq \epsilon_A \end{cases}$$

where σ *is a scalar such that* $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F = \epsilon_A$

Proof. Consider the constrained optimization problem for $\Delta\tilde{A}$:

$$\begin{aligned} \min_{\Delta\tilde{A}} \quad & \mathrm{Tr}(X^T(\tilde{A}+\Delta\tilde{A})^T(\tilde{A}+\Delta\tilde{A})X) \\ \text{s.t.} \quad & \|\Delta\tilde{A}\|_F^2 \leq \epsilon_A^2 \end{aligned} \tag{4.7}$$

The Lagrangian is

$$L(\Delta\tilde{A},\lambda) = \mathrm{Tr}(X^T(\tilde{A}+\Delta\tilde{A})^T(\tilde{A}+\Delta A)X) - \lambda(\|\Delta\tilde{A}\|_F^2 - \epsilon_A^2) \tag{4.8}$$

From matrix calculus, we know that

$$\begin{aligned} \frac{\partial \mathrm{Tr}(X^T(\tilde{A}+\Delta\tilde{A})^T(\tilde{A}+\Delta\tilde{A})X)}{\partial\Delta\tilde{A}} &= \frac{\partial \mathrm{Tr}(2X^T\tilde{A}^T\Delta\tilde{A}X)}{\partial\Delta\tilde{A}} + \frac{\partial \mathrm{Tr}(X^T\Delta\tilde{A}^T\Delta\tilde{A}X)}{\partial\Delta\tilde{A}} \\ &= \frac{\partial \mathrm{Tr}(2XX^T\tilde{A}^T\Delta\tilde{A})}{\partial\Delta\tilde{A}} + \frac{\partial \mathrm{Tr}(XX^T\Delta\tilde{A}^T\Delta\tilde{A})}{\partial\Delta\tilde{A}} \\ &= 2\tilde{A}XX^T + 2\Delta\tilde{A}XX^T \end{aligned} \tag{4.9}$$

The optimality condition $\nabla L(\Delta\tilde{A},\lambda) = 0$ is

$$\begin{cases} \tilde{A}XX^T + \Delta\tilde{A}XX^T + \lambda\Delta\tilde{A} = 0 \\ \|\Delta\tilde{A}\|_F^2 \leq \epsilon_A^2 \\ \lambda \geq 0 \\ \lambda(\|\Delta\tilde{A}\|_F^2 - \epsilon_A^2) = 0 \end{cases} \tag{4.10}$$

Assume that $\lambda = 0$, (4.10) becomes:

$$\begin{cases} \tilde{A}XX^T + \Delta\tilde{A}XX^T = 0 \\ \|\Delta\tilde{A}\|_F^2 \leq \epsilon_A^2 \end{cases} \tag{4.11}$$

The minimum norm solution for $\tilde{A}XX^T + \Delta\tilde{A}XX^T = 0$ is $\Delta\tilde{A} = -\tilde{A}X(X^TX)^{-1}X^T$
In other words, when $\|\tilde{A}X(X^TX)^{-1}X^T\|_F \leq \epsilon_A$, the solution to (4.10) is:

$$\begin{cases} \Delta\tilde{A} = -\tilde{A}X(X^TX)^{-1}X^T \\ \lambda = 0 \end{cases} \tag{4.12}$$

In this case, the optimal value of (4.6) is

$$\begin{aligned} &\mathrm{Tr}(X^T(\tilde{A} - \tilde{A}X(X^TX)^{-1}X^T)^T(\tilde{A} - \tilde{A}X(X^TX)^{-1}X^T)X) \\ =&\mathrm{Tr}((\tilde{A}X - \tilde{A}X(X^TX)^{-1}X^TX)^T(\tilde{A}X - \tilde{A}X(X^TX)^{-1}X^TX)) \\ =&\mathrm{Tr}((\tilde{A}X - \tilde{A}X)^T(\tilde{A}X - \tilde{A}X)) \\ =&0 \end{aligned} \tag{4.13}$$

When $\|\tilde{A}X(X^TX)^{-1}X^T\|_F > \epsilon_A$, the solution of (4.11) doe not exist, which means that $\lambda > 0$
Assume that $\lambda > 0$, (4.10) becomes:

$$\begin{cases} \tilde{A}XX^T + \Delta\tilde{A}XX^T + \lambda\Delta\tilde{A} = 0 \\ \|\Delta\tilde{A}\|_F = \epsilon_A \end{cases} \tag{4.14}$$

The solution for $\Delta\tilde{A}$ is $\Delta\tilde{A} = -\tilde{A}XX^T(\lambda I + XX^T)^{-1} = -\tilde{A}X(\lambda I + X^TX)^{-1}X^T$, where λ is a scalar such that $\|\tilde{A}X(\lambda I + X^TX)^{-1}X^T\|_F = \epsilon_A$
Note that such $\lambda > 0$ always exists because

$$\lim_{\lambda\to\infty} \|\tilde{A}X(\lambda I + X^TX)^{-1}X^T\|_F = 0 < \epsilon_A \tag{4.15}$$

and

$$\lim_{\lambda\to 0} \|\tilde{A}X(\lambda I + X^TX)^{-1}X^T\|_F = \|\tilde{A}X(X^TX)^{-1}X^T\|_F > \epsilon_A \tag{4.16}$$

In other words, when $\|\tilde{A}X(X^TX)^{-1}X^T\|_F > \epsilon_A$, the solution to (4.10) is:

$$\begin{cases} \Delta\tilde{A} = -\tilde{A}X(\lambda I + X^TX)^{-1}X^T \\ \|\tilde{A}X(\lambda I + X^TX)^{-1}X^T\|_F = \epsilon_A \end{cases} \tag{4.17}$$

In this case, the optimal value of (4.6) is

$$\begin{aligned}
&\mathrm{Tr}(X^T(\tilde{A} - \tilde{A}X(\lambda I + X^TX)^{-1}X^T)^T(\tilde{A} - \tilde{A}X(\lambda I + X^TX)^{-1}X^T)X) \\
=&\mathrm{Tr}((\tilde{A}X - \tilde{A}X(\lambda I + X^TX)^{-1}X^TX)^T(\tilde{A}X - AX(\lambda I + X^TX)^{-1}X^TX)) \\
=&\mathrm{Tr}((I - (\lambda I + X^TX)^{-1}X^TX)^TX^T\tilde{A}^T\tilde{A}X(I - (\lambda I + X^TX)^{-1}X^TX)) \\
=&\mathrm{Tr}((\lambda I + X^TX - X^TX)^T(\lambda I + X^TX)^{-T}X^T\tilde{A}^T\tilde{A}X(\lambda I + X^TX)^{-1}(\lambda I + X^TX - X^TX)) \\
=&\lambda^2\mathrm{Tr}((\lambda I + X^TX)^{-1}X^T\tilde{A}^T\tilde{A}X(\lambda I + X^TX)^{-1}) \\
=&\lambda^2\|\tilde{A}X(\lambda I + X^TX)^{-1}\|_F^2
\end{aligned} \tag{4.18}$$

Summarizing (4.13) and (4.18), we complete the proof. □

There are two cases depending on whether $\|\tilde{A}X(X^TX)^{-1}X^T\|_F \leq \epsilon_A$ or not. If we consider the outer maximization problem over X, since $\sigma^2\|\tilde{A}X(\sigma I + X^TX)^{-1}\|_F^2 \geq 0$, X will always be chosen such that $\|\tilde{A}X(X^TX)^{-1}X^T\|_F > \epsilon_A$ unless this is impossible. Now we study under what conditions there exists at least one solution for $\|\tilde{A}X(X^TX)^{-1}X^T\|_F > \epsilon_A$.

Lemma 6. $\max_X \|\tilde{A}X(X^TX)^{-1}X^T\|_F = \sqrt{\Sigma_{i=1}^k \sigma_i^2}$, *where* σ_i *is the* i^{th} *largest singular value of* $\tilde{A}$. *The maximum is achieved when the columns of* X *spans the leading* k *dimensional eigenspace of* $\tilde{A}^T\tilde{A}$.

Proof. Note that $X(X^TX)^{-1}X^T$ is invariant under linear transformations of columns. So without loss of generality, we assume that $X^TX = I$, then

$$\begin{aligned}
\|\tilde{A}X(X^TX)^{-1}X^T\|_F^2 &= \|\tilde{A}XX^T\|_F^2 \\
&= \mathrm{Tr}(XX^T\tilde{A}^T\tilde{A}XX^T) \\
&= \mathrm{Tr}(X^T\tilde{A}^T\tilde{A}XX^TX) \\
&= \mathrm{Tr}(X^T\tilde{A}^T\tilde{A}X)
\end{aligned} \tag{4.19}$$

According to (2.12), this is maximized when X consists of the first k eigenvectors of $\tilde{A}^T\tilde{A}$ and the maximal value is $\Sigma_{i=1}^k \lambda_i$, where λ_i is the i^{th} eigenvalue of $\tilde{A}^T\tilde{A}$ Therefore $\max_X \|\tilde{A}X(X^TX)^{-1}X^T\|_F = \sqrt{\Sigma_{i=1}^k \lambda_i} = \sqrt{\Sigma_{i=1}^k \sigma_i^2}$ □

The lemma says that if $\epsilon_A \geq \sqrt{\Sigma_{i=1}^k \sigma_i^2}$, we cannot find any X such that the objective value of (4.6) is nonzero. In other words, when ϵ_A is too large compared to $\tilde{A}$, it makes no sense to model the problem as worst-case robust optimization. In the following study, we assume that $\epsilon_A < \sqrt{\Sigma_{i=1}^k \sigma_i^2}$.

Now the problem becomes

$$\begin{aligned} \max_{X} \quad & \sigma(X)^2 \|\tilde{A}X(\sigma(X)I + X^TX)^{-1}\|_F^2 \\ \text{s.t.} \quad & X^TBX = I \end{aligned} \tag{4.20}$$

where $\sigma(X)$ is an implicit function defined by $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F = \epsilon_A$

This is a highly non-convex optimization problem. We will solve for this problem in the next section.

4.2 Optimality Condition

In this section, we derive the optimality conditions for the two non-convex optimization problems mentioned in the previous section.

Formulation 1

Consider the optimization problem:

$$\begin{aligned} \max_{X} \quad & \mathrm{Tr}(X^TAX) - \epsilon_A\|X^TX\|_F \\ \text{s.t.} \quad & X^TBX = I \end{aligned} \tag{4.21}$$

The Lagrangian is

$$L(X, \Lambda) = \mathrm{Tr}(X^TAX) - \epsilon_A\|X^TX\|_F - \mathrm{Tr}(\Lambda^T(X^TBX - I)) \tag{4.22}$$

The optimality condition $\nabla L(X, \Lambda) = 0$ is

$$\begin{cases} AX - \epsilon_A \dfrac{XX^TX}{\|XX^T\|_F} = BX\Lambda \\ X^TBX = I \end{cases} \tag{4.23}$$

To the best of our knowledge, this is an unsolved non-linear optimization problem that requires further study. In this paper, we stop here for this specific formulation and turn to the other formulation which we successfully found a way to solve.

Formulation 2

Consider the optimization problem:

$$\begin{aligned} \max_{X} \quad & \sigma(X)^2 \|\tilde{A}X(\sigma(X)I + X^TX)^{-1}\|_F^2 \\ \text{s.t.} \quad & X^TBX = I \end{aligned} \tag{4.24}$$

where $\sigma(X)$ is an implicit function defined by $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F = \epsilon_A$
It would be difficult to derive the optimality condition for X directly from this representation of the objective function. Let's consider another representation of the objective function.

$$\begin{aligned} \max_{X} \quad & \text{Tr}(X^T(\tilde{A} + D(X))^T(\tilde{A} + D(X))X) \\ \text{s.t.} \quad & X^TBX = I \end{aligned} \tag{4.25}$$

where $D(X)$ is an implicit matrix function defined by $D(X) = -\tilde{A}X(\sigma I + X^TX)^{-1}X^T$ and $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F = \epsilon_A$
The Lagrangian is

$$L(X, \Lambda) = \text{Tr}(X^T(\tilde{A} + D(X))^T(\tilde{A} + D(X))X) - \text{Tr}(\Lambda^T(X^TBX - I)) \tag{4.26}$$

Let $F(X) = G(X, D(X)) = \text{Tr}(X^T(\tilde{A} + D(X))^T(\tilde{A} + D(X))X)$, then

$$\begin{aligned} \frac{dF(X)}{dX} &= \frac{\partial G(X, D(X))}{\partial X} + \frac{\partial G(X, D(X))}{\partial D(X)} \times \frac{dD(X)}{dX} \\ &= 2(\tilde{A} + D(X))^T(\tilde{A} + D(X))X + \mathbf{0} \times \frac{dD(X)}{dX} \\ &= 2(\tilde{A} + D(X))^T(\tilde{A} + D(X))X \end{aligned} \tag{4.27}$$

where $\times$ denotes tensor multiplication
Then the optimality condition $\nabla L(X, \Lambda) = 0$ is

$$\begin{cases} (\tilde{A} - \tilde{A}X(\sigma I + X^TX)^{-1}X^T)^T(\tilde{A} - \tilde{A}X(\sigma I + X^TX)^{-1}X^T)X = BX\Lambda \\ \|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F = \epsilon_A \\ X^TBX = I \end{cases} \tag{4.28}$$

This is a non-linear system of equations of (σ, X). There might be multiple solutions and we are looking for the solution (σ^*, X^*) such that $\text{Tr}(\Lambda)$ is maximized.
Before diving into algorithms that solve (4.28), we need to simplify the first equation

of (4.10).
Using the identity of

$$X(\sigma I + X^T X)^{-1} = (\sigma I + XX^T)^{-1} X \tag{4.29}$$

and

$$I - (\sigma I + X^T X)^{-1} X^T X = \sigma(\sigma I + X^T X)^{-1} \tag{4.30}$$

We have

$$\begin{aligned}
&(\tilde{A} - \tilde{A}X(\sigma I + X^T X)^{-1} X^T)^T (\tilde{A} - \tilde{A}X(\sigma I + X^T X)^{-1} X^T) X \\
=&(I - X(\sigma I + X^T X)^{-1} X^T)^T \tilde{A}^T (\tilde{A}X - \tilde{A}X(\sigma I + X^T X)^{-1} X^T X) \\
=&(I - X(\sigma I + X^T X)^{-1} X^T)^T \tilde{A}^T \tilde{A}X(I - (\sigma I + X^T X)^{-1} X^T X) \\
=&(I - (\sigma I + XX^T)^{-1} XX^T)^T \tilde{A}^T \tilde{A}X(I - (\sigma I + X^T X)^{-1} X^T X) \\
=&\sigma(\sigma I + XX^T)^{-1} \tilde{A}^T \tilde{A}X\sigma(\sigma I + X^T X)^{-1} \\
=&\sigma^2(\sigma I + XX^T)^{-1} \tilde{A}^T \tilde{A}X(\sigma I + X^T X)^{-1}
\end{aligned} \tag{4.31}$$

Then the first equation of (4.28) is equivalent to

$$\sigma^2(\sigma I + XX^T)^{-1} \tilde{A}^T \tilde{A}X(\sigma I + X^T X)^{-1} = BX\Lambda \tag{4.32}$$

or

$$\sigma^2 \tilde{A}^T \tilde{A}X = (\sigma I + XX^T)BX\Lambda(\sigma I + X^T X) \tag{4.33}$$

Since $X^T BX = I$, this can be further simplified to

$$\sigma^2 \tilde{A}^T \tilde{A}X = (\sigma B + I)X\Lambda X^T(\sigma B + I)X \tag{4.34}$$

Now the optimality condition 4.28 becomes

$$\begin{cases} \sigma^2 \tilde{A}^T \tilde{A}X - (\sigma B + I)X\Lambda X^T(\sigma B + I)X = 0 \\ \|\tilde{A}X(\sigma I + X^T X)^{-1} X^T\|_F = \epsilon_A \\ X^T BX = I \end{cases} \tag{4.35}$$

4.3 The Subproblem

In this section, we propose an algorithm to solve for (4.35). We first solve the subproblem where σ is given, and then analyze the main problem of solving for (4.35).

The equation (4.35) is highly non-convex and difficult to solve directly. However, once σ is fixed, solving for $X(\sigma)$ based on the following turns out to be a tractable problem of numerical linear algebra.

$$\begin{cases} \sigma^2 \tilde{A}^T \tilde{A} X - (\sigma B + I) X \Lambda X^T (\sigma B + I) X = 0 \\ X^T B X = I \end{cases} \tag{4.36}$$

Let $\tilde{B}^T \tilde{B} = B$ and $Y = \tilde{B} X$, we have

$$\begin{cases} (\tilde{A}\tilde{B}^{-1})^T (\tilde{A}\tilde{B}^{-1}) Y - (I + \frac{1}{\sigma} \tilde{B}^{-T} \tilde{B}^{-1}) Y \Lambda Y^T (I + \frac{1}{\sigma} \tilde{B}^{-T} \tilde{B}^{-1}) Y = 0 \\ Y^T Y = I \end{cases} \tag{4.37}$$

Let $\Omega(Y) = \Lambda Y^T (I + \frac{1}{\sigma} \tilde{B}^{-T} \tilde{B}^{-1}) Y$, we have

$$\begin{cases} (\tilde{A}\tilde{B}^{-1})^T (\tilde{A}\tilde{B}^{-1}) Y - (I + \frac{1}{\sigma} \tilde{B}^{-T} \tilde{B}^{-1}) Y \Omega(Y) = 0 \\ Y^T Y = I \end{cases} \tag{4.38}$$

This is a non-linear system of equations of Y. However, the first equation is very similar to the generalized eigenvalue problem in the sense that the column spaces of Y is restricted to the eigensubspace of the generalized eigenvalue problem

$$(\tilde{A}\tilde{B}^{-1})^T (\tilde{A}\tilde{B}^{-1}) Y = (I + \frac{1}{\sigma} \tilde{B}^{-T} \tilde{B}^{-1}) Y \tag{4.39}$$

This is because $\Omega(Y)$ is multiplied by the right hand side of $(I + \frac{1}{\sigma} \tilde{B}^{-T} \tilde{B}^{-1}) Y$ and would not change the column space spanned by $(I + \frac{1}{\sigma} \tilde{B}^{-T} \tilde{B}^{-1}) Y$, which can be viewed as a 'scalar' matrix.

To describe the feasible linear subspaces Y that satisfies

$$(\tilde{A}\tilde{B}^{-1})^T (\tilde{A}\tilde{B}^{-1}) Y - (I + \frac{1}{\sigma} \tilde{B}^{-T} \tilde{B}^{-1}) Y \Omega(Y) = 0, \tag{4.40}$$

we can simultaneously diagonalize the two matrices:

$$R^T (\tilde{A}\tilde{B}^{-1})^T (\tilde{A}\tilde{B}^{-1}) R = \begin{pmatrix} c_1^2 & & \\ & \ddots & \\ & & c_n^2 \end{pmatrix}, \quad R^T (I + \tfrac{1}{\sigma} \tilde{B}^{-T} \tilde{B}^{-1}) R = \begin{pmatrix} s_1^2 & & \\ & \ddots & \\ & & s_n^2 \end{pmatrix}$$

where $R = (\mathbf{r}_1 \cdots \mathbf{r}_n)$ is an invertible matrix, and we can scale $c_i^2 + s_i^2 = 1$ to impose uniqueness up to diagonal permutation, which is a version of Cosine-sine(CS) decomposition.
Substituting $(\tilde{A}\tilde{B}^{-1})^T(\tilde{A}\tilde{B}^{-1})$ and $(I + \frac{1}{\sigma}\tilde{B}^{-T}\tilde{B}^{-1})$, we have

$$R^{-T}C^2R^{-1}Y - R^{-T}S^2R^{-1}Y\Omega(Y) = 0 \tag{4.41}$$

Simplifying the equation, we have

$$(S^{-2}C^2)(R^{-1}Y) = (R^{-1}Y)\Omega(Y) \tag{4.42}$$

Since $\Omega(Y)$ would not change the column spaces spanned by $(R^{-1}Y)$, $(R^{-1}Y)$ must be an eigenspace of $(S^{-2}C^2)$, which is a diagonal matrix. While an eigenspace of a diagonal matrix consists of $\mathbf{e}_i = (0, \ldots, 1, \ldots, 0)^T$, this means that there are at most k non-zero rows in $(R^{-1}Y)$. In other words, any feasible linear subspace Y consists of r columns of R. Since the decomposition is unique up to diagonal permutation, without loss of generality, we can write:

$$Y = R\begin{pmatrix} U_1 \\ 0 \end{pmatrix} = \begin{pmatrix} R_1 & R_2 \end{pmatrix}\begin{pmatrix} U_1 \\ 0 \end{pmatrix} = R_1U_1 \tag{4.43}$$

where U_1 is a $k \times k$ square matrix.

So in general, we have $\binom{n}{k}$ solutions. One benefit of using such a decomposition is that we can now write $\mathrm{Tr}(\Lambda)$ and $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F$ as a function of C, S and R

Lemma 7. *$Tr(\Lambda) = \Sigma_{i=1}^k \frac{c_i^2}{s_i^4}\|\boldsymbol{r}_i\|^2$*

Proof. According to (4.42) and (4.43), we have

$$\begin{pmatrix} S_1^{-2}C_1^2 & 0 \\ 0 & S_2^{-2}C_2^2 \end{pmatrix}\begin{pmatrix} U_1 \\ 0 \end{pmatrix} = \begin{pmatrix} U_1 \\ 0 \end{pmatrix}\Omega \tag{4.44}$$

Therefore

$$\Omega = U_1^{-1}S_1^{-2}C_1^2U_1 \tag{4.45}$$

On the other hand,

$$\begin{aligned}\Omega &= \Lambda Y^T(I + \frac{1}{\sigma}\tilde{B}^{-T}\tilde{B}^{-1})Y \\ &= \Lambda(R\begin{pmatrix}U_1\\0\end{pmatrix})^T(R^{-T}S^2R^{-1})(R\begin{pmatrix}U_1\\0\end{pmatrix}) \\ &= \Lambda\begin{pmatrix}U_1^T & 0\end{pmatrix}\begin{pmatrix}S_1^2 & 0\\0 & S_2^2\end{pmatrix}\begin{pmatrix}U_1\\0\end{pmatrix} \\ &= \Lambda U_1^T S_1^2 U_1\end{aligned} \tag{4.46}$$

Combing (4.45) and (4.46), we have

$$\begin{aligned}\Lambda &= U_1^{-1}S_1^{-2}C_1^2U_1(U_1^TS_1^2U_1)^{-1} \\ &= U_1^{-1}S_1^{-2}C_1^2U_1U_1^{-1}S_1^{-2}U_1^{-T} \\ &= U_1^{-1}S_1^{-4}C_1^2U_1^{-T}\end{aligned} \tag{4.47}$$

Since $U_1^TR_1^TR_1U_1 = (R_1U_1)^T(R_1U_1) = Y^TY = I$, we have

$$U_1^{-T}U_1^{-1} = R_1^TR_1 \tag{4.48}$$

Therefore

$$\begin{aligned}\mathrm{Tr}(\Lambda) &= \mathrm{Tr}(U_1^{-1}S_1^{-4}C_1^2U_1^{-T}) \\ &= \mathrm{Tr}(S_1^{-4}C_1^2U_1^{-T}U_1^{-1}) \\ &= \mathrm{Tr}(S_1^{-4}C_1^2R_1^TR_1) \\ &= \mathrm{Tr}(R_1S_1^{-4}C_1^2R_1^T) \\ &= \Sigma_{i=1}^k\frac{c_i^2}{s_i^4}\|\mathbf{r}_i\|^2\end{aligned} \tag{4.49}$$

□

Lemma 8. $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F^2 = \frac{1}{\sigma}(\Sigma_{i=1}^k\frac{c_i^2}{s_i^2} - \Sigma_{i=1}^k\frac{c_i^2}{s_i^4}\|\boldsymbol{r}_i\|_2^2)$

Proof.

$$\begin{aligned}&\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F^2 \\ =&\mathrm{Tr}((\tilde{A}X(\sigma I + X^TX)^{-1}X^T)^T\tilde{A}X(\sigma I + X^TX)^{-1}X^T) \\ =&\mathrm{Tr}(X(\sigma I + X^TX)^{-1}X^T\tilde{A}^T\tilde{A}X(\sigma I + X^TX)^{-1}X^T) \\ =&\mathrm{Tr}((\sigma I + X^TX)^{-1}X^T\tilde{A}^T\tilde{A}X(\sigma I + X^TX)^{-1}X^TX) \\ =&\mathrm{Tr}((\sigma I + X^TX)^{-1}X^T\tilde{A}^T\tilde{A}X(\sigma I + X^TX)^{-1}(\sigma I + X^TX - \sigma I)) \\ =&\mathrm{Tr}((\sigma I + X^TX)^{-1}X^T\tilde{A}^T\tilde{A}X) - \sigma\mathrm{Tr}((\sigma I + X^TX)^{-1}X^T\tilde{A}^T\tilde{A}X(\sigma I + X^TX)^{-1})\end{aligned} \tag{4.50}$$

For the second term,

$$\begin{aligned}
&\sigma\mathrm{Tr}((\sigma I + X^T X)^{-1} X^T \tilde{A}^T \tilde{A} X (\sigma I + X^T X)^{-1}) \\
=&\sigma \|\tilde{A} X (\sigma I + X^T X)^{-1}\|_F^2 \\
=&\frac{1}{\sigma}(\sigma^2 \|\tilde{A} X (\sigma I + X^T X)^{-1}\|_F^2) \\
=&\frac{1}{\sigma}\mathrm{Tr}(\Lambda) \\
=&\frac{1}{\sigma}(\Sigma_{i=1}^k \frac{c_i^2}{s_i^4}\|\mathbf{r}_i\|^2)
\end{aligned} \tag{4.51}$$

For the first term,

$$\begin{aligned}
&\mathrm{Tr}((\sigma I + X^T X)^{-1} X^T \tilde{A}^T \tilde{A} X) \\
=&\mathrm{Tr}((\sigma I + Y^T \tilde{B}^{-T} \tilde{B}^{-1} Y)^{-1} Y^T \tilde{B}^{-T} \tilde{A}^T \tilde{A} \tilde{B}^{-1} Y) \\
=&\mathrm{Tr}((\sigma Y^T Y + Y^T \tilde{B}^{-T} \tilde{B}^{-1} Y)^{-1} (R \begin{pmatrix} U_1 \\ 0 \end{pmatrix})^T (\tilde{A}\tilde{B}^{-1})^T (\tilde{A}\tilde{B}^{-1}) R \begin{pmatrix} U_1 \\ 0 \end{pmatrix}) \\
=&\mathrm{Tr}((Y^T(\sigma I + \tilde{B}^{-T}\tilde{B}^{-1})Y)^{-1} \begin{pmatrix} U_1^T & 0 \end{pmatrix} (R^T (\tilde{A}\tilde{B}^{-1})^T (\tilde{A}\tilde{B}^{-1}) R) \begin{pmatrix} U_1 \\ 0 \end{pmatrix}) \\
=&\mathrm{Tr}(((R \begin{pmatrix} U_1 \\ 0 \end{pmatrix})^T (\sigma I + \tilde{B}^{-T}\tilde{B}^{-1}) R \begin{pmatrix} U_1 \\ 0 \end{pmatrix})^{-1} \begin{pmatrix} U_1^T & 0 \end{pmatrix} \begin{pmatrix} C_1^2 & 0 \\ 0 & C_2^2 \end{pmatrix} \begin{pmatrix} U_1 \\ 0 \end{pmatrix}) \\
=&\frac{1}{\sigma}\mathrm{Tr}(((\begin{pmatrix} U_1^T & 0 \end{pmatrix} (R^T (I + \frac{1}{\sigma} G^{-T} \tilde{B}^{-1}) R) \begin{pmatrix} U_1 \\ 0 \end{pmatrix})^{-1} U_1^T C_1^2 U_1) \\
=&\frac{1}{\sigma}\mathrm{Tr}(((\begin{pmatrix} U_1^T & 0 \end{pmatrix} \begin{pmatrix} S_1^2 & 0 \\ 0 & S_2^2 \end{pmatrix} \begin{pmatrix} U_1 \\ 0 \end{pmatrix})^{-1} U_1^T C_1^2 U_1) \\
=&\frac{1}{\sigma}\mathrm{Tr}((U_1^T S_1^2 U_1)^{-1} U_1^T C_1^2 U_1) \\
=&\frac{1}{\sigma}\mathrm{Tr}(U_1^{-1} S_1^{-2} U_1^{-T} U_1^T C_1^2 U_1) \\
=&\frac{1}{\sigma}\mathrm{Tr}(U_1^{-1} S_1^{-2} C_1^2 U_1) \\
=&\frac{1}{\sigma}\mathrm{Tr}(S_1^{-2} C_1^2 U_1 U_1^{-1}) \\
=&\frac{1}{\sigma}\mathrm{Tr}(S_1^{-2} C_1^2) \\
=&\frac{1}{\sigma}(\Sigma_{i=1}^k \frac{c_i^2}{s_i^2})
\end{aligned} \tag{4.52}$$

Combining (4.51) and (4.52), we have

$$\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F^2 = \frac{1}{\sigma}(\Sigma_{i=1}^k \frac{c_i^2}{s_i^2} - \Sigma_{i=1}^k \frac{c_i^2}{s_i^4}\|\mathbf{r}_i\|_2^2) \tag{4.53}$$

□

From these two lemma, we not only represent $\mathrm{Tr}(\Lambda)$ and $\|\tilde{A}X(\sigma I+X^TX)^{-1}X^T\|_F$ as a function of C, S and R, but also understand that these two quantities are separable among n indices. Furthurmore, they have some nice properties under specific conditions.

Lemma 9. *When $\tilde{A}^T\tilde{A}B = B\tilde{A}^T\tilde{A}$, $f_i(\sigma) = \frac{c_i^2(\sigma)}{s_i^4(\sigma)}\|\boldsymbol{r}_i(\sigma)\|^2$ is an increasing function*

Proof. When $\tilde{A}^T\tilde{A}B = B\tilde{A}^T\tilde{A}$,

$$\begin{aligned}(A\tilde{B}^{-1})^T(\tilde{A}\tilde{B}^{-1})(\tilde{B}^{-T}\tilde{B}^{-1}) &= \tilde{B}^{-T}A^TA\tilde{B}^{-1}\tilde{B}^{-T}\tilde{B}^{-1}\\ &= \tilde{B}^{-T}\tilde{A}^T\tilde{A}B^{-1}\tilde{B}^{-1}\\ &= \tilde{B}^{-T}B^{-1}\tilde{A}^T\tilde{A}\tilde{B}^{-1}\\ &= \tilde{B}^{-T}\tilde{B}^{-1}\tilde{B}^{-T}\tilde{A}^T\tilde{A}\tilde{B}^{-1}\\ &= \tilde{B}^{-T}\tilde{B}^{-1}(\tilde{A}\tilde{B}^{-1})^T(A\tilde{B}^{-1})\\ &= (\tilde{B}^{-T}\tilde{B}^{-1})(\tilde{A}\tilde{B}^{-1})^T(A\tilde{B}^{-1})\end{aligned} \tag{4.54}$$

So $(\tilde{A}\tilde{B}^{-1})^T(\tilde{A}\tilde{B}^{-1})$ and $\tilde{B}^{-T}\tilde{B}^{-1}$ commute, then there exists an orthogonal matrix P such that

$$P^T(\tilde{A}\tilde{B}^{-1})^T(\tilde{A}\tilde{B}^{-1})P = \begin{pmatrix} a_1^2 & & \\ & \ddots & \\ & & a_n^2 \end{pmatrix}, \quad P^T(\tilde{B}^{-T}\tilde{B}^{-1})P = \begin{pmatrix} b_1^2 & & \\ & \ddots & \\ & & b_n^2 \end{pmatrix}$$

Therefore the CS decomposition has analytic solution of σ:

$$R^T(\tilde{A}\tilde{B}^{-1})^T(\tilde{A}\tilde{B}^{-1})R = \begin{pmatrix} c_1^2 & & \\ & \ddots & \\ & & c_n^2 \end{pmatrix}, \quad R^T(I + \tfrac{1}{\sigma}\tilde{B}^{-T}\tilde{B}^{-1})R = \begin{pmatrix} s_1^2 & & \\ & \ddots & \\ & & s_n^2 \end{pmatrix}$$

where $c_i^2 = a_i^2w_i^2$, $s_i^2 = (1 + \frac{b_i^2}{\sigma})w_i^2$, $w_i = \frac{1}{\sqrt{a_i^2+1+\frac{b_i^2}{\sigma}}}$, $R = PW$, $W = \mathrm{diag}(w_i)$

Then we can write $f_i(\sigma)$ as an explicit function of σ:

$$f_i(\sigma) = \frac{c_i^2(\sigma)}{s_i^4(\sigma)}\|\mathbf{r}_i(\sigma)\|^2 = \frac{a_i^2w_i^2}{(1+\frac{b_i^2}{\sigma})^2w_i^4}w_i^2 = \frac{a_i^2}{(1+\frac{b_i^2}{\sigma})^2} \tag{4.55}$$

This is an increasing function of σ □

Lemma 10. *When* $\tilde{A}^T\tilde{A}B = B\tilde{A}^T\tilde{A}$, $g_i(\sigma) = \frac{1}{\sigma}(\frac{c_i^2(\sigma)}{s_i^2(\sigma)} - \frac{c_i^2(\sigma)}{s_i^4(\sigma)}\|\boldsymbol{r}_i(\sigma)\|^2)$ *is a decreasing function*

Proof. According to (4.54), when $\tilde{A}^T\tilde{A}B = B\tilde{A}^T\tilde{A}$, $(\tilde{A}\tilde{B}^{-1})^T(\tilde{A}\tilde{B}^{-1})$ and $\tilde{B}^{-T}\tilde{B}^{-1}$ commute, and then the CS decomposition has analytic solution of σ:

$$R^T(\tilde{A}\tilde{B}^{-1})^T(\tilde{A}\tilde{B}^{-1})R = \begin{pmatrix} c_1^2 & & \\ & \ddots & \\ & & c_n^2 \end{pmatrix}, \quad R^T(I + \tfrac{1}{\sigma}\tilde{B}^{-T}\tilde{B}^{-1})R = \begin{pmatrix} s_1^2 & & \\ & \ddots & \\ & & s_n^2 \end{pmatrix}$$

where $c_i^2 = a_i^2 w_i^2$, $s_i^2 = (1 + \frac{b_i^2}{\sigma})w_i^2$, $w_i = \frac{1}{\sqrt{a_i^2+1+\frac{b_i^2}{\sigma}}}$, $R = PW$, $W = \mathrm{diag}(w_i)$

Then we can write $g_i(\sigma)$ as an explicit function of σ:

$$\begin{aligned} g_i(\sigma) &= \frac{1}{\sigma}(\frac{c_i^2(\sigma)}{s_i^2(\sigma)} - \frac{c_i^2(\sigma)}{s_i^4(\sigma)}\|\mathbf{r}_i(\sigma)\|^2) \\ &= \frac{1}{\sigma}(\frac{a_i^2 w_i^2}{(1+\frac{b_i^2}{\sigma})w_i^2} - \frac{a_i^2 w_i^2}{(1+\frac{b_i^2}{\sigma})^2 w_i^4} w_i^2) \\ &= \frac{1}{\sigma}(\frac{a_i^2}{(1+\frac{b_i^2}{\sigma})} - \frac{a_i^2}{(1+\frac{b_i^2}{\sigma})^2}) \\ &= \frac{1}{\sigma^2}\frac{a_i^2 b_i^2}{(1+\frac{b_i^2}{\sigma})^2} \\ &= \frac{a_i^2 b_i^2}{(\sigma + b_i^2)^2} \end{aligned} \tag{4.56}$$

This is a decreasing function of σ □

Corollary 1. *When* $\tilde{A}^T\tilde{A}B = B\tilde{A}^T\tilde{A}$, *among all* $\binom{n}{k}$ *solutions to (4.36), the* j^{th} *largest* $\|\tilde{A}X(\sigma)(\sigma I + X(\sigma)^T X(\sigma))^{-1}X(\sigma)^T\|_F^2$ *is a decreasing function over* σ *for any* j

This corollary says that if we view all solutions given σ as $\binom{n}{k}$ solution paths, then along these paths, $\|\tilde{A}X(\sigma)(\sigma I + X(\sigma)^T X(\sigma))^{-1}X(\sigma)^T\|_F^2$ is a deceasing vector with respect to σ. This is useful when we use bisection to find the root $\|\tilde{A}X(\sigma)(\sigma I + X(\sigma)^T X(\sigma))^{-1}X(\sigma)^T\|_F^2 = \epsilon_A$ in the next section. Before diving into the next section, we conjecture that these nice properties hold for general $\tilde{A}$ and B:

Conjecture 1. *$f_i(\sigma) = \frac{c_i^2(\sigma)}{s_i^4(\sigma)}\|\boldsymbol{r}_i(\sigma)\|^2$ is an increasing function*

Conjecture 2. *$g_i(\sigma) = \frac{1}{\sigma}(\frac{c_i^2(\sigma)}{s_i^2(\sigma)} - \frac{c_i^2(\sigma)}{s_i^4(\sigma)}\|\boldsymbol{r}_i(\sigma)\|^2)$ is a decreasing function*

4.4 The Brute-force Algorithm

Now let's go back to the optimality condition (4.35):

$$\begin{cases} \sigma^2\tilde{A}^T\tilde{A}X - (\sigma B + I)X\Lambda X^T(\sigma B + I)X = 0 \\ \|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F = \epsilon_A \\ X^TBX = I \end{cases}$$

In general, there might be multiple solutions and we are looking for the solution (σ^*, X^*) such that $\mathrm{Tr}(\Lambda)$ is maximized. Given σ, we've already seen that there are exactly $\binom{n}{k}$ solutions to the following sub problem:

$$\begin{cases} \sigma^2\tilde{A}^T\tilde{A}X - (\sigma B + I)X\Lambda X^T(\sigma B + I)X = 0 \\ X^TBX = I \end{cases} \tag{4.57}$$

Under the condition that $\tilde{A}^T\tilde{A}$ and B commute with each other, or under the conjectures 1,2, any of these solutions has the property that $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F$ is decreasing over σ. This means that there would be at most $\binom{n}{k}$ solutions to the optimality condition (4.35). This motivates the following bisection algorithm for finding all of the roots of (4.35) and thus the optimal solution to the problem (4.6):

Step 1: Given σ,j, solve for X with j^{th} largest $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F$

Step 2: If $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F > \epsilon_A$, set $\sigma_{new} > \sigma$, otherwise, set $\sigma_{new} < \sigma$

Step 3: Iterate until the root X_j^* is found and record $\mathrm{Tr}(\Lambda_j^*)$

Step 4: Enumerate j and find the maximal $\mathrm{Tr}(\Lambda_j^*)$

This is an algorithm that tries to solve for all stationary points, i.e. solutions to the necessary optimality conditions (4.35), and find the one with largest objective. The time complexity is polynomial in n but exponential in k, which is $O(n^{2k}\log(\frac{1}{\epsilon}))$.

This is not practical when n or k is large. Whether there exists an algorithm that has substantially lower time complexity remains an open question. Here we proposed a heuristic algorithm that performs reasonably well in practice.

4.5 The Heuristic Algorithm

We propose to solve for the solution $(\tilde{\sigma}, \tilde{X})$ such that σ is maximized. This way, instead of checking every single path while iterating over σ, we only need to check the specific path with maximal $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F$. The algorithm to find the solution $(\tilde{\sigma}, \tilde{X})$ such that σ is maximized is as follows:

Step 1: Given σ, solve for X with the largest $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F$

Step 2: If $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F > \epsilon_A$, set $\sigma_{new} > \sigma$, otherwise, set $\sigma_{new} < \sigma$

Step 3: Iterate until the stop criterion is satisfied

Step 1 is summarized in algorithm 3. In step 1, we solve for the CS decomposition (4.40) given σ. There would be $\binom{n}{k}$ solutions but we would be only interested in whether there are any solution X such that $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F > \epsilon_A$. This can be done in $O(n)$ time since $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F$ is separable among indices, as discussed in the previous section. Consider $g_i(\sigma) = \frac{1}{\sigma}(\frac{c_i^2(\sigma)}{s_i^2(\sigma)} - \frac{c_i^2(\sigma)}{s_i^4(\sigma)}\|\mathbf{r}_i(\sigma)\|^2)$ and then the indices corresponding to the largest k values would give us the column space of X, which is exactly the X with the largest $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F$. If $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F > \epsilon_A$, the equation (4.35) would have another solution pair $(\hat{\sigma}, \hat{X})$ such that $\hat{\sigma} > \sigma$, and this means that the maximal $\tilde{\sigma}$ among all feasible solutions would be greater than or equal to σ. Therefore in step 2, we test another σ_{new} that is larger than σ. On the other hand, if none of the solution satisfies $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F > \epsilon_A$, this means that that the maximal $\tilde{\sigma}$ among all feasible solutions would be less than σ. Therefore in step 2, we test another σ_{new} that is less than σ. This way, after every iteration, σ_{new} gets closer and closer to the optimal value $\tilde{\sigma}$ and we can use some stopping criteria to terminate the iteration.

There are many ways to update σ_{new}. Here we proposed the bisection algorithm 4 to update σ_{new}.

Algorithm 3 Path Selection

Input: matrix $\tilde{A}, \tilde{B}$, param ϵ_A, σ, k
$\mathbf{c}, \mathbf{s}, R = \text{decomposition}((\tilde{A}\tilde{B}^{-1})^T(\tilde{A}\tilde{B}^{-1}), I + \frac{1}{\sigma}\tilde{B}^{-T}\tilde{B}^{-1})$
$a = \text{argsort}(\frac{c_i^2(\sigma)}{s_i^2(\sigma)} - \frac{c_i^2(\sigma)}{s_i^4(\sigma)}\|\mathbf{r}_i(\sigma)\|^2)$
$idx = a[:k]$
$Y = R[:, idx]$
$Q, \_ = QR(Y)$
$X = B^{-1}Y$
Output: matrix X

Algorithm 4 Bisection

Input: matrix $\tilde{A}, B$, param ϵ_A, ub
$lb, ub, i = 0, ub, 0$
while $i <$ MAXITER:
  $\sigma = (lb + ub)/2$
  $X = \text{Path\_Selection}(\tilde{A}, \sqrt{B}, \sigma)$
  if $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F > \epsilon_A$
    $lb = \sigma$
  else:
    $ub = \sigma$
  $i = i + 1$
$\sigma^* = (lb + ub)/2$
$X^* = \text{Path\_Selection}(\tilde{A}, \sqrt{B}, \sigma)$
Output: optimal solution X^*

4.6 Properties

In this section, we discuss the properties of our proposed heuristic algorithm 4.

First, the algorithm 4 always converges to a stationary point (σ, X), i.e. one solution to the optimality condition (4.35)

Theorem 3. *Let σ_j be the output of the j^{th} step from the bisection algorithm 4, then there exist $\tilde{\sigma}$ such that $\sigma_j \to \tilde{\sigma}$ with linear rate of convergence.*

Proof. The algorithm 4 is in fact a bisection algorithm finding the root $\tilde{\sigma}$ of:

$$g(\sigma) = \|\tilde{A}X(\sigma)(\sigma I + X(\sigma)^TX(\sigma))^{-1}X(\sigma)^T\|_F^2 - \epsilon_A^2 = 0 \tag{4.58}$$

where $X(\sigma)$ is determined by the algorithm 3.
According to lemma 8, we know that

$$g(\sigma) = \max_{\mathcal{I}} \Sigma_{i\in\mathcal{I}} \frac{1}{\sigma}(\frac{c_i^2}{s_i^2} - \frac{c_i^2}{s_i^4}\|\mathbf{r}_i\|_2^2) \tag{4.59}$$

where the index set $\mathcal{I}$ consists of all possible k indices
So $g(\sigma)$ is a continuous function of σ since $c_i, s_i, \mathbf{r}_i$ from CS decomposition are all continuous function of σ.
Also, since $X^TBX = I$, elements of X are upper bounded, therefore

$$\lim_{\sigma\to\infty} g(\sigma) = \lim_{\sigma\to\infty} \|\tilde{A}X(\sigma)(\sigma I + X(\sigma)^T X(\sigma))^{-1}X(\sigma)^T\|_F^2 - \epsilon_A^2 = -\epsilon_A^2 < 0 \tag{4.60}$$

Also, we have

$$\lim_{\sigma\to 0} g(\sigma) = \lim_{\sigma\to 0} \|\tilde{A}X(\sigma)(\sigma I + X(\sigma)^T X(\sigma))^{-1}X(\sigma)^T\|_F^2 - \epsilon_A^2 = \Sigma_{i=1}^k \sigma_i^2 - \epsilon_A^2 > 0 \tag{4.61}$$

where σ_i is the i^{th} largest singular value of $\tilde{A}$. Note that the assumption $\Sigma_{i=1}^k \sigma_i^2 > \epsilon_A^2$ is discussed in 6.
The previous two limits make sure that the bisection works and has linear rate of convergence. □

The next question is whether this is the optimal solution to the original problem (4.6). It turns out that in general, the answer is no. This is why this is only a heuristic algorithm. However, under certain conditions, this solution is the solution $(\tilde{\sigma}, \tilde{X})$ such that σ is maximized.

Lemma 11. *If $\tilde{A}^T\tilde{A}B = B\tilde{A}^T\tilde{A}$, or conjecture 2 holds, then $(\tilde{\sigma}, \tilde{X})$ obtained from the algorithm 4 is the solution to (4.35) such that σ is maximized.*

Proof. According to lemma 10 and conjecture 2, in this case,

$$g(\sigma) = \|\tilde{A}X(\sigma)(\sigma I + X(\sigma)^T X(\sigma))^{-1}X(\sigma)^T\|_F^2 - \epsilon_A^2 \tag{4.62}$$

is decreasing. Suppose there exist another solution to (4.35) $(\hat{\sigma}, \hat{X})$ such that $\hat{\sigma} > \tilde{\sigma}$, then

$$0 = g(\tilde{\sigma}) > g(\hat{\sigma}) > \|\tilde{A}\hat{X}(\sigma I + \hat{X}^T\hat{X})^{-1}\hat{X}^T\|_F^2 - \epsilon_A^2 > 0 \tag{4.63}$$

which is a contradiction. Therefore, $(\tilde{\sigma}, \hat{X})$ is the solution to (4.35) such that σ is maximized. □

The benefit of obtaining the maximal σ solution is that under certain conditions, we can easily have an upper bound on the optimal value to (4.6). Given $\tilde{\sigma}$, we solve for (4.36) and choose X such that $\mathrm{Tr}(\Lambda)$ is maximized. We will show in Theorem 4 that this would be an upper bound of the optimal value of the original problem (4.6). To be more specific, we define the following function:

$$h(\sigma) = \max_{\mathcal{I}} \Sigma_{i\in\mathcal{I}} \frac{c_i^2(\sigma)}{s_i^4(\sigma)} \|\mathbf{r}_i(\sigma)\|^2 \tag{4.64}$$

where the index set $\mathcal{I}$ consists of all possible k indices

Theorem 4. *Let (σ^*, X^*) be the optimal solution of (4.6), $f(X^*)$ be the optimal value, and $(\tilde{\sigma}, \tilde{X})$ be the solution to (4.35) such that σ is maximized. If $\tilde{A}^T\tilde{A}B = B\tilde{A}^T\tilde{A}$, or conjecture 1 holds, then $f(\tilde{X}) \leq f(X^*) \leq h(\tilde{\sigma})$*

Proof. Since $\tilde{X}$ is a valid solution to the optimality condition (4.35) and thus a valid solution to the original problem (4.6), the function value $f(X)$ serves as a natural lower bound of the optimal value $f(X^*)$.
On the other hand, let $\mathcal{I}^*$ be the index set chosen by the optimal solution X^* at σ^*,

$$h(\tilde{\sigma}) \geq \Sigma_{i\in\mathcal{I}^*} \frac{c_i^2(\tilde{\sigma})}{s_i^4(\tilde{\sigma})} \|\mathbf{r}_i(\tilde{\sigma})\|^2 \geq \Sigma_{i\in\mathcal{I}^*} \frac{c_i^2(\sigma^*)}{s_i^4(\sigma^*)} \|\mathbf{r}_i(\sigma^*\|^2 = f(X^*) \tag{4.65}$$

The first inequality comes from the definition of h, and the second inequality comes from lemma 9 and the fact that $\tilde{\sigma} \geq \sigma$. □

This theorem suggests an improved version of algorithm 4:

Step 1: Given σ, solve for X with the largest $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F$

Step 2: If $\|\tilde{A}X(\sigma I + X^TX)^{-1}X^T\|_F > \epsilon_A$, set $\sigma_{new} > \sigma$, otherwise, set $\sigma_{new} < \sigma$

Step 3: Iterate until the stop criterion is satisfied and yield the solution pair $(\tilde{\sigma}, \tilde{X})$

Step 4: Calculate $h(\tilde{\sigma})$ and yield the interval for optimal value $[f(\tilde{X}), h(\tilde{\sigma})]$

This improved version of the algorithm tells the user to what extend the solution is close the optimal. Surprisingly, in practice, $f(\tilde{X}) = h(\tilde{\sigma})$ for most of the time when $\tilde{A}$ and B are set to be random Gaussian matrices. This means under certain conditions, our proposed algorithm can usually find the global optimal solution and we believe that this is a practical heuristic algorithm.

Chapter 5

Robust Adaptive Beamforming

5.1 Introduction

It is widely recognized that adaptive beamforming methods' performance significantly declines when the desired signal is included in the training data, even with minor discrepancies in the knowledge of the desired signal covariance matrix. Such mismatches between the assumed and actual source covariance matrices can arise due to factors like antenna element displacement, changing environments, or imperfections in the propagation medium, among others. The primary objective of any robust adaptive beamforming (RAB) approach is to ensure resilience against these types of mismatches.

The majority of RAB methods have been designed for point source signals when the rank of the desired signal covariance matrix is equal to one. In many practical scenarios, such as incoherently scattered signal sources or sources with fluctuating (randomly distorted) wave-fronts, the source covariance matrix's rank is greater than one. While the RAB methods in [1] offer excellent robustness against mismatches based on the point source assumption, they are not ideally suited for cases when the rank of the desired signal covariance matrix exceeds one.

The general-rank signal model RAB, which explicitly models error mismatches, was developed in [41] using the worst-case performance optimization principle. Although the RAB in [41] has a straightforward closed-form solution, it is excessively conservative, as the worst-case correlation matrix for the desired signal could be indefinite or even negative definite. Consequently, less conservative approaches were introduced in [15],[42], incorporating an additional positive semi-definite (PSD) con-

straint on the worst-case signal covariance matrix. The main drawback of the RAB methods in [15],[42] is that they only provide a suboptimal solution, potentially leaving a significant gap to the global optimal solution. For instance, the RAB in [15] iteratively finds a suboptimal solution, but there is no guarantee of convergence. A closed-form approximate suboptimal solution is proposed in [42]; however, this solution might also be far from the global optimal one. These limitations prompt the exploration of new, efficient strategies to solve the aforementioned non-convex problem in a globally optimal manner.

In this chapter, we will first describe the system model in detail in section 2, and then formulate the robust optimization problem in section 3. Numerical experiments are shown in section 4. Finally, we explore the potential use of our general-rank algorithm in the rarely explored field of multi-rank beamforming in section 5.

5.2 System Model

At a given time instant t, the linear antenna array, consisting of M omni-directional antenna elements, receives a narrow band signal. This can be described as:

$$x(t) = s(t) + i(t) + n(t) \tag{5.1}$$

where $s(t), i(t)$, and $n(t)$ are the $M \times 1$ vectors of the desired signal, interference, and noise, respectively. They are assumed to be statistically independent. The output of the beamformer at time t is:

$$y(t) = w^H x(t) \tag{5.2}$$

where w is the $M \times 1$ complex beamforming vector of the antenna array. The challenge of beamforming can be characterized as determining the optimal beamforming vector w, which enhances the beamformer output's signal-to-interference-plus-noise ratio (SINR) to its maximum potential. This is expressed as:

$$\text{SINR} = \frac{w^H R_s w}{w^H R_{i+n} w} \tag{5.3}$$

where R_s and R_{i+n} are defined as:

$$R_s = E[s(t)s^H(t)] \tag{5.4}$$

$$R_{i+n} = E[(i(t) + n(t)(i(t) + n(t))^H] \tag{5.5}$$

Based on the characteristics of the desired signal source, its corresponding covariance matrix can possess a varying rank, that is, $1 \leq \text{rank}(R_s) \leq M$. In numerous practical applications, such as scenarios involving incoherently scattered signal sources or signals with randomly fluctuating wave-fronts, the desired signal covariance matrix's rank exceeds one. The unique instance where the rank is equal to one occurs in the case of a point source.

5.3 Problem Formulation

Adaptive beamforming solves the following optimization problem:

$$\max_{w} \quad \frac{w^H R_s w}{w^H R_{i+n} w} \tag{5.6}$$

or

$$\max_{w} \quad \frac{w^H R_s w}{w^H R w} \tag{5.7}$$

where w is the vector of beamformer weights, $R = R_s + R_{i+n}$ is the mixed covariance matrix, and R_s and R_{i+n} are the signal and interference-plus-noise covariance matrices respectively. This is a generalized eigenvalue problem and can be solved analytically. However, in practical scenarios, neither R_s or R_{i+n} is exactly known. To provide robustness against matrix mismatches Δ_1 and Δ_2 in the two matrices, researchers proposed to solve the following worst case robust optimization problem:

$$\max_{w} \min_{||\Delta_1|| \leq \eta, ||\Delta_2|| \leq \epsilon} \quad \frac{||(Q + \Delta_1)w||_2^2}{w^H (R + \Delta_2) w} \tag{5.8}$$

where $R_s = Q^H Q$
This is equivalent of solving for the following optimization problem:

$$\max_{w} \quad \frac{(||Qw||_2 - \eta ||w||_2)^2}{w^H (R + \epsilon I) w} \tag{5.9}$$

This problem is non-convex and researchers have been developing algorithms for solving convex approximations to this problem. However, we can solve this non convex optimization problem directly without any convex approximation.

In the context of robust adaptive beamforming, the signal covariance matrix R_s depends on the signal angular power density [30]:

$$R_s = \sigma_s^2 \int_{-\pi/2}^{\pi/2} \xi(\theta) a(\theta) a^H(\theta) d\theta \tag{5.10}$$

where σ_s is the desired signal power, $\xi(\theta)$ is the normalized signal angular power density function, and $a(\theta)$ is the steering vector towards direction θ.

In practice, σ_s is a scalar that has no impact on the optimal w, and $a(\theta)$ is an exactly known vector function that only depends on the set up of antennas. For example, for a uniform linear array of N antennas,

$$a_k(\theta) = e^{i\pi k \sin(\theta)}, \quad k = 1, 2, \ldots, N \tag{5.11}$$

However, the normalized angular power density is unknown, and people have to presume such a function based on their prior information about the source of signal and the environment of transmission.

As for the denominator, we have an empirical covariance matrix $\hat{R}$ and this is used to approximate the mixed covariance matrix $R \approx \hat{R}$.

5.4 Numerical Simulations

In our experiment, we consider a uniform linear array (ULA) of 10 omni-directional antenna elements with the inter-element spacing of half wave length. The power of noise is defined to be 0 dB and the interference-to-noise ratio (INR) is set to be 20 dB. The interferer is locally incoherently scattered with uniform angle power density with central angle of $-30°$ and angular spread of $10°$. The desired signal is locally incoherently scattered with Gaussian angle power density with central angle of $30°$ and angular spread of $4°$. This is not exactly known and there are four main types of data mismatch. We might overestimate or underestimate the central angle or the angular spread. Therefore here we consider four presumed signal angle power density:

(a) Gaussian angle power density with central angle of $32°$ and angular spread of $5°$.

(b) Gaussian angle power density with central angle of $32°$ and angular spread of $3°$.

(c) Gaussian angle power density with central angle of $28°$ and angular spread of $5°$.

(d) Gaussian angle power density with central angle of $28°$ and angular spread of $3°$.

Furthermore, we assume that the mismatch distance or upper bound is roughly known and the mismatch parameters are set to be $\eta = \sqrt{||\tilde{R}_s - R_s||}$ and $\epsilon = ||\hat{R} - R||$.

We compared our proposed method with three other methods and the statistically optimal method:

Optimal This solves (5.7)

$$\max_{w \neq 0} \quad \frac{w^H R_s w}{w^H R w}$$

with exact R_s and R, which means that there is no data mismatch. Note that this is not practical and is only used for reference.

MVDR Vanilla MVDR directly optimizes (5.7)

$$\max_{w \neq 0} \quad \frac{w^H \hat{R}_s w}{w^H \hat{R} w}$$

with presumed $\hat{R}_s$ and empirical $\hat{R}$.

Robust MVDR without PSD constraint Robust adaptive beamforming without the positive semi-definite constraint [41] solves (3.2):

$$\max_{w \neq 0} \min_{\|\Delta A\| \leq \epsilon_A, \|\Delta B\| \leq \epsilon_B} \quad \frac{w^H (\hat{R}_s + \Delta A) w}{w^H (\hat{R} + \Delta B) w}$$

Approximate Robust MVDR with PSD constraint Robust adaptive beamforming with the positive semi-definite constraint [42] approximates (5.8) with closed form solution:

$$w_2^{opt} = \mathcal{P}((\hat{R} + \epsilon_B I)^{-1} (\hat{R}_s - 2\sqrt{\lambda_{max}(\hat{R}_s)} \eta I + \epsilon_A^2 I))$$

The performance of all of these methods in the four cases (a),(b),(c), and (d) are shown in figures 5.1,5.2,5.3, and 5.4, respectively. The signal-to-interference-plus-noise-ratio (SINR) are compared versus different scale of signal-to-noise-ratio (SNR).

The result shows that in case (a) and (c), our proposed method and the method of [42] are better among the three methods. In case (b) and (d), our proposed method and the method of [41] are better among the three. Therefore, our proposed method performs constantly well in the considered types of data uncertainty.

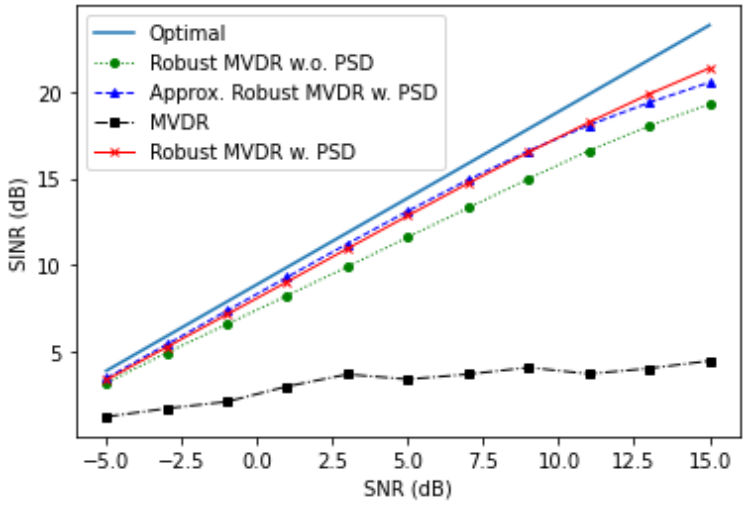

Figure 5.1: Output SINR versus different SNR. Case (a): the presumed signal covariance matrix is Gaussian angle power density with central angle of 32° and angular spread of 5°.

5.5 Multi-rank Beamforming

In our experiment, we consider a uniform linear array (ULA) of 10 omni-directional antenna elements with the inter-element spacing of half wave length. The power of noise is defined to be 0 dB and the interference-to-noise ratio (INR) is set to be 20 dB. The interferer is locally incoherently scattered with uniform angle power density with central angle of $-30°$ and angular spread of 10°. The desired signal is assumed to be consist of two parts. Each part of the signal is locally incoherently scattered with Gaussian angle power density. The first part is with central angle of 30° and angular spread of 4° and the second part is with central angle of 60° and angular spread of 4°. In a word, the angular power density of the desired signal is assumed to follow a Gaussian mixture model. The two distributions are not exactly known and we assume that presumed signal of the second part is with central angle of 57° and angular spread of 3°, and we consider four types of data mismatch of the first part of the signal:

(a) Gaussian angle power density with central angle of 32° and angular spread of 5°.

(b) Gaussian angle power density with central angle of 32° and angular spread of 3°.

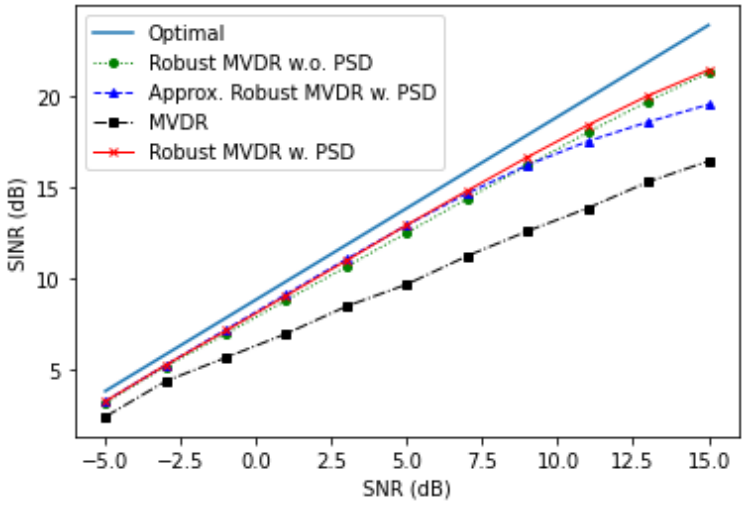

Figure 5.2: Output SINR versus different SNR. Case (b): the presumed signal covariance matrix is Gaussian angle power density with central angle of 32° and angular spread of 3°.

(c) Gaussian angle power density with central angle of 28° and angular spread of 5°.

(d) Gaussian angle power density with central angle of 28° and angular spread of 3°.

Furthermore, we assume that the mismatch distance or upper bound is roughly known and the mismatch parameters are set to be $\eta = \sqrt{||\tilde{R}_s - R_s||}$ and $\epsilon = ||\hat{R} - R||$.

To the best of our knowledge, there is no competing algorithms for the problem at hand. Therefore we only compared our proposed method with the vanilla MVDR, which optimizes (5.7) directly with presumed $\hat{R}_s$ and empirical $\hat{R}$.

The performance of all of these methods in the four cases (a),(b),(c), and (d) are shown in Figures 5.5,5.6,5.7, and 5.8, respectively. The signal-to-interference-plus-noise-ratio (SINR) are compared versus different scale of signal-to-noise-ratio (SNR).

The result shows that our proposed method performs constantly well in the considered types of data uncertainty.

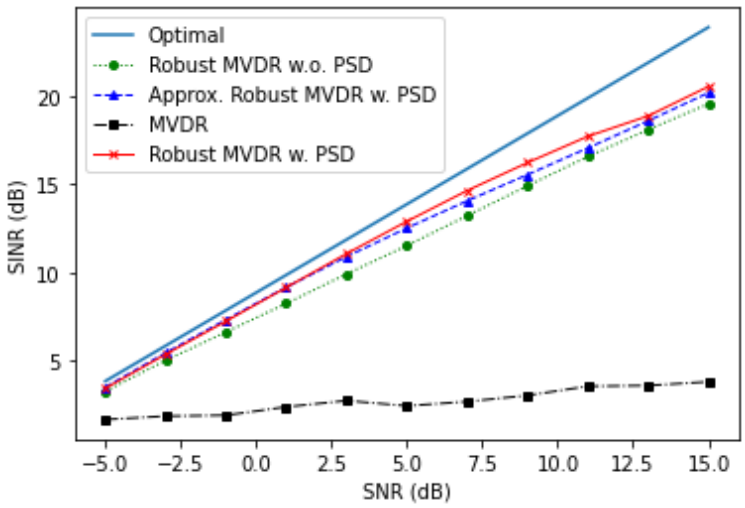

Figure 5.3: Output SINR versus different SNR. Case (c): the presumed signal covariance matrix is Gaussian angle power density with central angle of 28° and angular spread of 5°.

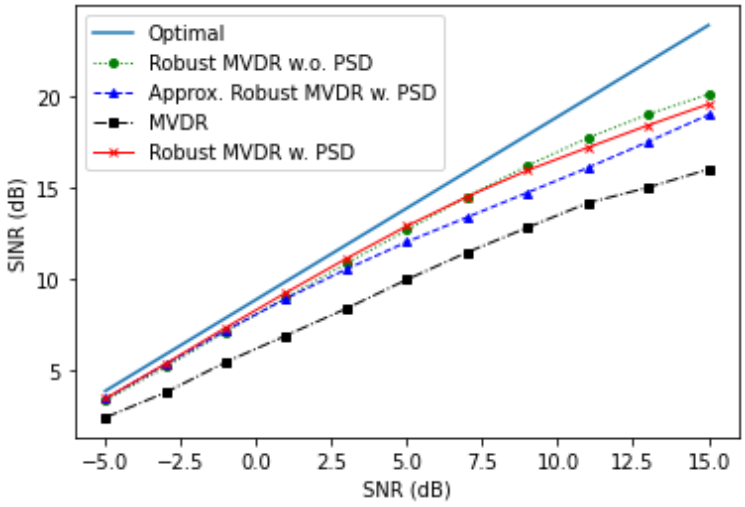

Figure 5.4: Output SINR versus different SNR. Case (d): the presumed signal covariance matrix is Gaussian angle power density with central angle of 28° and angular spread of 3°.

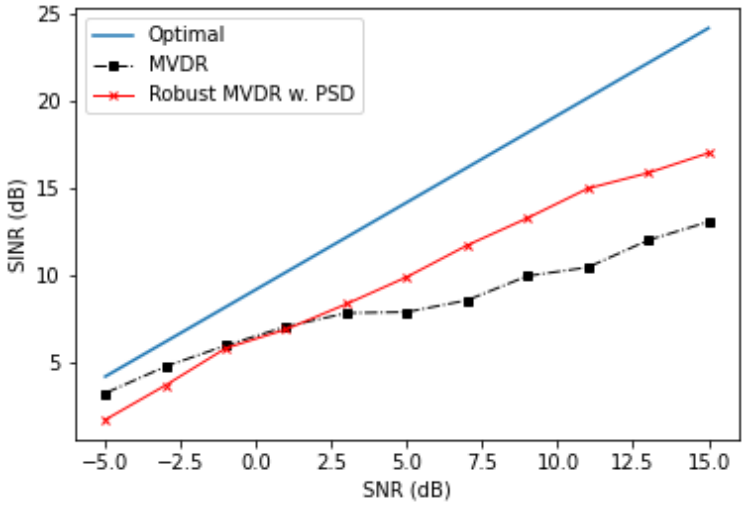

Figure 5.5: Output SINR versus different SNR. Case (a): the presumed signal covariance matrix is Gaussian angle power density with central angle of 32° and angular spread of 5°.

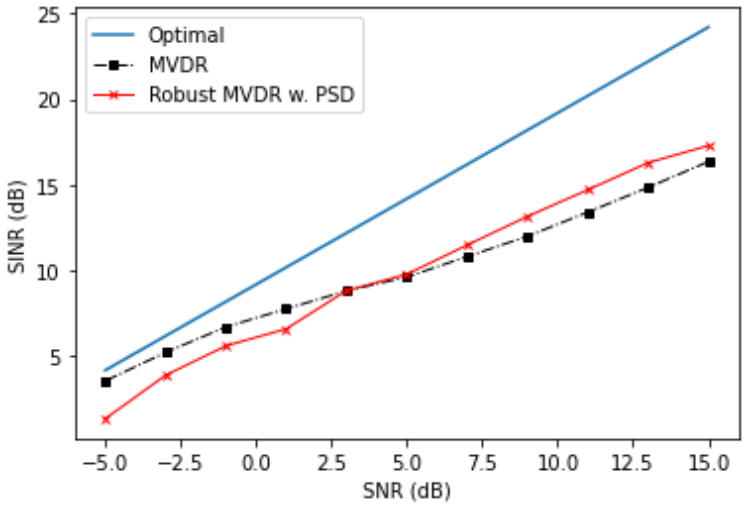

Figure 5.6: Output SINR versus different SNR. Case (b): the presumed signal covariance matrix is Gaussian angle power density with central angle of 32° and angular spread of 3°.

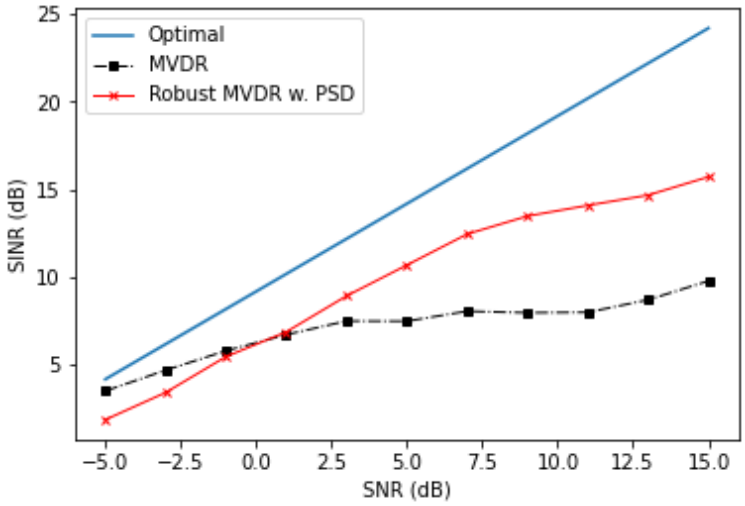

Figure 5.7: Output SINR versus different SNR. Case (c): the presumed signal covariance matrix is Gaussian angle power density with central angle of 28° and angular spread of 5°.

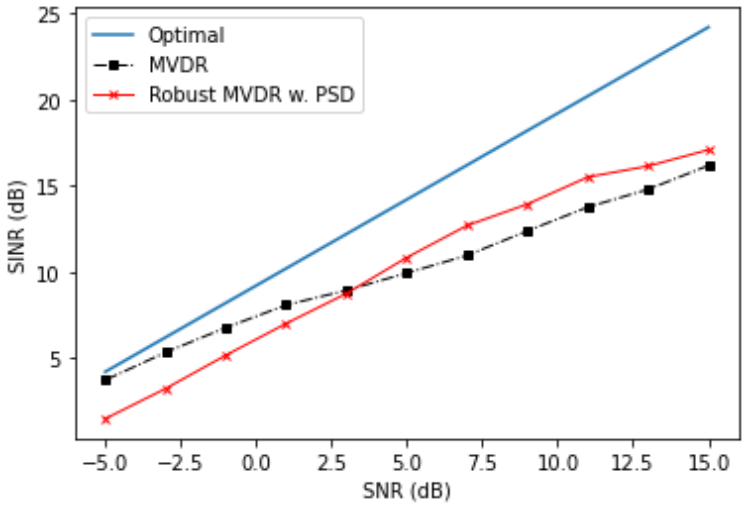

Figure 5.8: Output SINR versus different SNR. Case (d): the presumed signal covariance matrix is Gaussian angle power density with central angle of 28° and angular spread of 3°.

Chapter 6

Semi-supervised Linear Discriminant Analysis

6.1 Introduction

Dimension reduction in machine learning is crucial for several reasons, including noise reduction, computational efficiency, visualization, overfitting prevention, and improved interpretability. By simplifying high-dimensional data, we can focus on essential features, save computational resources, and better visualize patterns and relationships. Moreover, reducing dimensions helps prevent overfitting by creating simpler models and enhances the interpretability of the results, making it easier to understand and explain the model's behavior. Overall, dimension reduction is an indispensable tool for optimizing the performance and efficiency of machine learning models. Notable examples of dimension reduction techniques include Principle commponent analysis(PCA) and Linear discriminant analysis(LDA).

LDA is a supervised method. It looks for directions where data points from different groups are far apart while keeping data points in the same group close together. When we have label information, like in classification tasks, LDA can perform much better than PCA[36]. However, when there aren't enough training samples compared to the number of dimensions, we might not accurately estimate each group's mean vector and covariance matrix. In this situation, we can't guarantee good results on test samples. A possible solution is learning from both labeled and unlabeled data, which is also called semi-supervised learning. This approach makes sense because, in real life, we often have only some labeled data and a large amount of unlabeled data.

In this chapter, we aim at dimensionality reduction in the semi-supervised case. It is well-known that supervised LDA is a generalized eigenvalue problem. We proposed to use worst-case robust optimization of generalized eigenvalue problem as a way to solve the semi supervised LDA problem. We first formulate the mathematical optimization problem in section 2. Experiment setting is then described in section 3. We then discuss numerical simulations of the rank-one case and the general-rank case separately in section 4 and section 5, respectively.

6.2 Problem Formulation

In Fisher discriminant analysis [37], or linear discriminant analysis [35], people solve the following optimization problem:

$$\max_{w} \quad \frac{w^T S_b w}{w^T S w} \tag{6.1}$$

or in the general-rank case:

$$\max_{W} \quad \mathrm{Tr}((W^T S W)^{-1} W^T S_b W) \tag{6.2}$$

where $w \in \mathbf{R}^{n\times 1}$ and $W \in \mathbf{R}^{n\times k}$ are the projection directions and projection subspaces, respectively. S_b is the between-class scatters, and S is the within-class scatters. They are defined as follows:

$$S_b = \frac{1}{C}\Sigma_{i=1}^{C}(\mu_i - \mu)(\mu_i - \mu)^T \tag{6.3}$$

$$S = \Sigma_{i=1}^{C} S_i \tag{6.4}$$

where $\mu_i \in \mathbf{R}^{n\times 1}$ and $S_i \in \mathbf{R}^{n\times n}$ are means and covariance matrices of the i^{th} class of points.

Classical linear discriminant analysis requires to know the class each point belongs to. In the absence of labels, if we know that each class has the same number of points and we have a rough estimate of the center of each class, we can reformulate the optimization problem and perform semi-supervised linear discriminant analysis. The idea is to consider the covariance matrix of all data points instead of class-specific covariance:

$$S_w = \mathrm{Cov}(x) = E[(x-\mu)(x-\mu)^T] \tag{6.5}$$

Then

$$\begin{aligned} S_w &= E[(x-\mu)(x-\mu)^T] \\ &= E[E[(x-\mu)(x-\mu)^T | x \in C_i]] \\ &= \frac{1}{C}\Sigma_{i=1}^{C} E[(x_i-\mu)(x_i-\mu)^T] \\ &= \frac{1}{C}\Sigma_{i=1}^{C} E[(x_i-\mu_i+(\mu_i-\mu))(x_i-\mu_i+(\mu_i-\mu))^T] \\ &= S_b + \frac{1}{C}S \end{aligned} \tag{6.6}$$

Therefore the optimization problem (6.1),(6.2) are equivalent to the following problems, respectively:

$$\max_{w} \quad \frac{w^T S_b w}{w^T S_w w} \tag{6.7}$$

$$\max_{W} \quad \mathrm{Tr}((W^T S_w W)^{-1} W^T S_b W) \tag{6.8}$$

The benefit of solving (6.7) and (6.8) instead of (6.1) and (6.2) is that in the absence of labels, it is usually very difficult to have prior information or rough estimation of S_i. In (6.7) or (6.8), we only need prior information of μ_i and S_w can be estimated using all unlabeled data points. The optimization problems (6.7) and (6.8) are generalized eigenvalue problems and can be solved analytically. However, neither S_b or S_w is exactly known. To provide robustness against matrix mismatches Δ_b and Δ_w in the two matrices, we can solve the following worst case robust optimization problem in the rank-one case:

$$\max_{w} \min_{||\Delta_b|| \le \epsilon_b, ||\Delta_w|| \le \epsilon_w} \quad \frac{||(Q+\Delta_b)w||_2^2}{w^T(S_w+\Delta_w)w} \tag{6.9}$$

where $S_b = Q^T Q$ and the following worst case robust optimization problem in the general-rank case:

$$\max_{W} \min_{||\Delta_b|| \le \epsilon_b} \quad \mathrm{Tr}((W^T S_w W)^{-1} W^T (Q+\Delta_b)^T (Q+\Delta_b) W) \tag{6.10}$$

where $S_b = Q^T Q$
This is equivalent to the following problem:

$$\begin{aligned} \max_{W} \min_{||\Delta_b|| \le \epsilon_b} \quad & \mathrm{Tr}(W^T (Q+\Delta_b)^T (Q+\Delta_b) W) \\ \text{s.t.} \quad & W^T S_w W = I \end{aligned} \tag{6.11}$$

Experiment Setting

In our experiment, we assume that the data follows a multivariate Gaussian mixture model:

$$p_i(x) \sim N(\mu_i, S_i), i = 1, 2, \dots K \tag{6.12}$$

We further assume that the dimension of data $M = 10$, the number of classes $K = 4$, and the number of data points for each class is $N = 500$. The covariance matrix S_w is approximated by the sample covariance matrix $\hat{S}_w$. Also, in our setting of semi-supervised linear discriminant analysis, we assume that we have inexact prior information about the means of these four classes $\hat{\mu}_i$. The directions of mismatch of the means are random and the scale of the mismatch is set to $||\mu_i - \hat{\mu}_i|| = 0.2||\mu_i - \mu||$, where μ is the mean of data in all classes. Note that the experiment result is not sensitive to the dimension M, the number of classes K, or the number of data N. Instead, it is sensitive to the degree of mixing of these four types of points. We used the intermixture value $IV = \frac{\sqrt{||S_i||}}{||\mu_i - \mu||}$ to quantify the degree of mixing. The true μ_i and S_i are randomly sampled in each simulation and we run 1000 independent simulations for each intermixture value.

6.3 Numerical Simulations: Rank-one Case

We compared our proposed method with three other methods and the statistically optimal method:

Optimal This solves (6.7)

$$\max_w \quad \frac{w^T S_b w}{w^T S_w w}$$

with exact S_b and S_w, which means that there is no data mismatch. Note that this is not practical and is only used for reference.

ULDA ULDA directly optimizes (6.7)

$$\max_w \quad \frac{w^T \hat{S}_b w}{w^T \hat{S}_w w}$$

with presumed $\hat{S}_b$ and empirical $\hat{S}_w$.

Robust ULDA without PSD constraint Robust adaptive beamforming without the positive semi-definite constraint [41] solves (3.2):

$$\max_{w\neq 0} \min_{\|\Delta A\|\leq\epsilon_A,\|\Delta B\|\leq\epsilon_B} \frac{w^T(\hat{S}_b+\Delta A)w}{w^T(\hat{S_w}+\Delta B)w}$$

Approximate Robust ULDA with PSD constraint Robust adaptive beamforming with the positive semi-definite constraint [42] approximates (5.8) with closed form solution:

$$w_2^{opt} = \mathcal{P}((\hat{S_w}+\epsilon_B I)^{-1}(\hat{S}_b - 2\sqrt{\lambda_{max}(\hat{S}_b)}\eta I + \epsilon_A^2 I))$$

The performance of all of these methods is shown in figure 6.1. The Fisher's metric, or the signal-to-noise-ratio SNR $= \frac{w^T S_b w}{w^T S w}$ are compared versus different intermixture value.

The result shows that in the low intermixture region, which corresponds to well-separated cases, our proposed method and the method of [42] are better among the four methods and are close to optimal value. While in the high intermixture region, the robust formulation cannot provide additional value. This also shows that in this specific context, the approximation [42] works as well as the true solution to the problem (6.9).

6.4 Numerical Simulations: General-rank Case

To the best of our knowledge, there is no competing algorithms for the problem at hand (6.11). Therefore we only compared our proposed method with the vanilla LDA, which optimizes (6.8) directly with presumed S_b and empirical $\hat{S}_w$.

The performance of the methods is shown in figure 6.2. The Fisher's metric, or the signal-to-noise-ratio SNR $= \mathrm{Tr}((W^T S_w W)^{-1} W^T S_b W)$ are compared versus different intermixture value.

The result shows that in the low intermixture region, which corresponds to well-separated cases, our proposed method is close to optimal and is significantly better than vanilla LDA. While in the high intermixture region, the robust formulation cannot provide significant additional value. This conclusion is similar to what we had in the rank-one case.

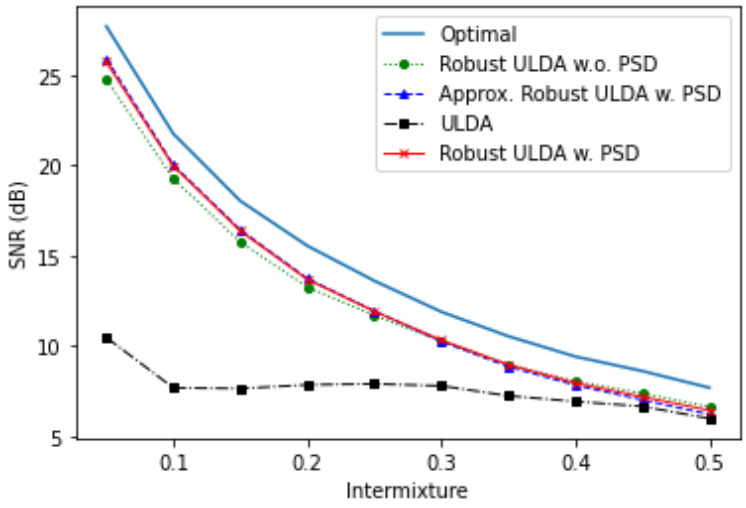

Figure 6.1: SNR versus intermixture of semi-supervised linear discriminant analysis.

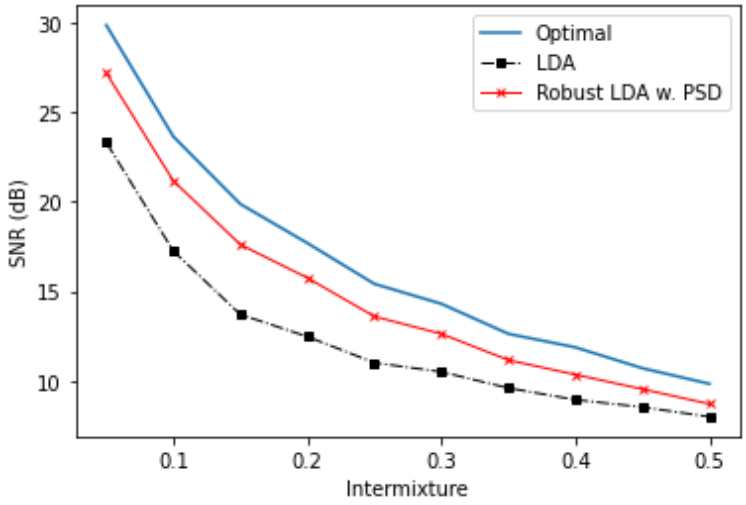

Figure 6.2: SNR versus intermixture of semi-supervised linear discriminant analysis.

Bibliography

[1] A.B.Gershman. "Robust adaptive beamforming in sensor arrays". In: *Journal of Electron Common* 53 (1999), pp. 305–314.

[2] A.B.Gershman, E.Nemeth, and J.F.Böhme. "Experimental performance of adaptive beamforming in a sonar environment with a towed array and moving interfering sources". In: *IEEE Transaction on Signal Processing* 48 (2000), pp. 246–250.

[3] A.B.Gershman, V.I.Turchin, and V.A.Zverev. "Experimental results of localization of moving underwater signal by adaptive beamforming". In: *IEEE Transaction on Signal Processing* 43 (1995), pp. 2249–2257.

[4] E. Anderson et al. *LAPACK Users' Guide*. Third. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1999. ISBN: 0-89871-447-8 (paperback).

[5] B.D.Carlson. "Covariance matrix estimation errors and diagonal loading in adaptive arrays". In: *IEEE Transactions on Aerospace and Electronic Systems* 24 (1988), pp. 397–401.

[6] B.N.Parlett. "The symmetric eigenvalue problem". In: *Classics in Applied Mathematics* 20 (1998).

[7] A. Ben-Tal, S. Boyd, and A. Nemirovski. "Extending Scope of Robust Optimization: Comprehensive Robust Counterparts of Uncertain Problems". In: (2005).

[8] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.

[9] A. Ben-Tal and A. Nemirovski. "Robust Convex Optimization". In: *Math. Oper. Res.* 23 (1998), pp. 769–805.

[10] A. Ben-Tal and A. Nemirovski. "Robust Solutions of Uncertain Linear Programs". In: *Operations Research* 25 (1999), pp. 1–13.

[11] A. Ben-Tal and A. Nemirovski. "Robust Truss Topology Design via Semidefinite Programming". In: *SIAM Journal on Optimization* 7 (1997), pp. 991–1016.

[12] A. Ben-Tal, A. Nemirovski, and C. Roos. "Robust solutions of uncertain quadratic and conic-quadratic problems". In: *SIAM Journal on Optimization* 13 (2002), pp. 535–560.

[13] S. Boyd and V. Balakrishnan. "A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its L∞-norm". In: *Systems Control Letters* 15 (1990), pp. 1–7.

[14] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[15] H. Chen and A. B. Gershman. "Robust adpative beamforming for general-rank signal models using positive semidefinite covariance constraints". In: *Proceedings of IEEE ICASSP* (2008), pp. 2341–2344.

[16] D.Cai, X.He, and J.Han. "Semi-supervised discriminant analysis". In: *ICCV* (2007).

[17] D.D.Feldman and L.J.Griffiths. "A projection approach to robust adaptive beamforming". In: *IEEE Transaction on Signal Processing* 42 (1994), pp. 867–876.

[18] E.Y.Gorodetskaya et al. "Deep-water acoustic coherence at long ranges: Theoretical prediction and effects on large-array signal processing". In: *IEEE Journal of Oceanic Engineering* 24 (1999), pp. 156–171.

[19] Ayman Elnashar. *Simplified Robust Adaptive Detection and Beamforming for Wireless Communications*. Wiley, 2018.

[20] L. El-Ghaoui and H. Lebret. "Robust Solutions to Least-Square Problems to Uncertain Data Matrices". In: *SIAM Journal on Matrix Analysis* 18 (1997), pp. 1035–1064.

[21] B. Ghojogh, F. Karray, and M. Crowley. "Eigenvalue and Generalized Eigenvalue Problems: Tutorial". In: *arXiv* (2022).

[22] H.Cox. "Resolving power and sensitivity to mismatch of optimum array processors". In: *The Journal of the Acoustical Society of America* 54 (1973), pp. 758–771.

[23] Y. Huang and S. A. Vorobyov. "An inner approximate algorithm for robust adaptive beamforming for general-rank signal model". In: *IEEE Signal Processing Letters* 25 (2018), pp. 1735–1739.

[24] J.Capon, R.J.Greenfield, and R.J.Kolker. "Multidimensional maximum-likelihood processing for a large aperture seismic array". In: *Proceedings of the IEEE* 55 (1967), pp. 192–211.

[25] J.H.Wilkinson. *The algebraic eigenvalue problem*. Vol. 662. Oxford Clarendon, 1965.

[26] J.L.Krolik. "The performance of matched-field beamformers with Mediterranean vertical array data". In: *IEEE Transaction on Signal Processing* 44 (1996), pp. 2605–2611.

[27] J.Riba, J.Goldberg, and G.Vazquez. "Robust beamforming for interference rejection in mobile communications". In: *IEEE Transaction on Signal Processing* 45 (1987), pp. 271–275.

[28] K.Fukunaga. *Introduction to statistical pattern recognition*. Academic Press Professional, 1990.

[29] K.L.Bell, Y.Ephraim, and H.L.Van Trees. "A Bayesian approach to robust adaptive beamforming". In: *IEEE Transaction on Signal Processing* 48 (2000), pp. 386–398.

[30] A. Khabbazibasmenj and S.A. Vorobyov. "Robust adaptive beamformingfor general-rank signal model with positive semi-definite constraint via POTDC". In: *IEEE Transactions on Signal Processing* 61 (2013), pp. 6103–6117.

[31] L.C.Godara. "Application of antenna arrays to mobile communications. II. Beam-forming and direction-of-arrival considerations". In: *Proceedings of the IEEE* 85 (1997), pp. 1195–1245.

[32] L.E.Brennan, J.D.Mallet, and I.S.Reed. "Adaptive arrays in airborne MTI radar". In: *IEEE Transaction on Antennas Propagation* 24 (1976), pp. 607–615.

[33] M.Sugiyama et al. "Semi-supervised local fisher discriminant analysis for dimensionality reduction". In: *Machine Learning* 78 (2010), pp. 35–61.

[34] M.Wax and Y.Anu. "Performance analysis of the minimum variance beamformer in the presence of steering vector errors". In: *IEEE Transaction on Signal Processing* 44 (1996), pp. 938–947.

[35] Geoffrey J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Interscience, 2004. ISBN: 978-0-471-69115-0.

[36] P.N.Belhumeur, J.P.Hepanha, and D.J.Kriegman. "Eigenfaces vs. fisherfaces: recognition using class specific linear projection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (1997), pp. 711–720.

[37] R.A.Fisher. "The Use of Multiple Measurements in Taxonomic Problems". In: *Annals of Eugenics* 7 (1936), pp. 179–188.

[38] R.A.Monzingo and T.W.Miller. *Introduction to Adaptive Arrays*. Wiley, 1980.

[39] R.O.Duda, P.E.Hart, and D.G.Stork. *Pattern Classification*. Wiley Interscience, 2000.

[40] Theodore S.Rappaport. *Smart Antennas: Adaptive Arrays, Algorithms, and Wireless Position Location*. 1998. ISBN: 978-0780348004.

[41] S. Shahbazpanahi et al. "Robust adaptive beamforming for general-rank signal models". In: *IEEE Transactions on Signal Processing* 51 (2003), pp. 2257–2269.

[42] C. Xing, S. Ma, and Y.-C. Wu. "On low complexity robust beamforming with positive semi-definite constraints". In: *IEEE Transactions on Signal Processing* 57 (2009), pp. 4942–4945.

[43] Y.Kameda and J.Ohga. "Adaptive microphone-array system for noise reduction". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 34 (1986), pp. 1391–1400.

[44] Y.Zhang and D.Y.Yeung. "Semi-supervised discriminant analysis via cccp". In: *ECML* (2008).