Robust Rayleigh quotient minimization and nonlinear eigenvalue problems

Zhaojun Bai[*], Ding Lu[†], and Bart Vandereycken[‡]

August 1, 2018

Abstract

We study the robust Rayleigh quotient optimization problem where the data matrices of the Rayleigh quotient are subject to uncertainties. We propose to solve such a problem by exploiting its characterization as a nonlinear eigenvalue problem with eigenvector nonlinearity (NEPv). For solving the NEPv, we show that a commonly used iterative method can be divergent due to a wrong ordering of the eigenvalues. Two strategies are introduced to address this issue: a spectral transformation based on nonlinear shifting and a reformulation using second-order derivatives. Numerical experiments for applications in robust generalized eigenvalue classification, robust common spatial pattern analysis, and robust linear discriminant analysis demonstrate the effectiveness of the proposed approaches.

1 Introduction

For a pair of symmetric matrices $A, B \in \mathbb{R}^{n\times n}$ with either $A \succ 0$ or $B \succ 0$ (positive definite), the Rayleigh quotient (RQ) minimization problem is to find the optimizers of

$$\min_{\substack{z\in\mathbb{R}^n\\ z\neq 0}} \frac{z^T A z}{z^T B z}. \tag{1}$$

It is well known that the optimal RQ corresponds to the smallest (real) eigenvalue of the generalized Hermitian eigenvalue problem $Az = \lambda Bz$. There exist many excellent methods for computing this eigenvalue by either computing the full eigenvalue decomposition or by employing large-scale eigenvalue solvers that target only the smallest eigenvalue and possibly a few others. We refer to [10, 3] and the references therein for an overview.

Eigenvalue problems and, in particular, RQ minimization problems, have numerous applications. Traditionally, they occur in the study of vibrations of mechanical structures but other applications in science and engineering including buckling, elasticity, control theory, and statistics. In quantum physics, eigenvalues and eigenvectors represent energy levels and orbitals of atoms and molecules. More recently, they are also a building block of many algorithms in data science and machine learning; see, e.g. [34, Chap. 2]. Of particular importance to the current paper are, for example, generalized eigenvalue classifiers [20], common spatial pattern analysis [7] and Fisher's

[*] Department of Computer Science and Department of Mathematics, University of California, Davis, CA 95616, USA. (bai@cs.ucdavis.edu)

[†] Department of Mathematics, University of Geneva, Geneva, Switzerland. (Ding.Lu@unige.ch)

[‡] Department of Mathematics, University of Geneva, Geneva, Switzerland. (Bart.Vandereychen@unige.ch)

linear discriminant analysis [8]. In §5 we will explain these examples in more detail, but in each example the motivation for solving the generalized eigenvalue problem comes from its relation to the RQ minimization (1).

1.1 Robust Rayleigh quotient minimization

Real-world applications typically use data that is not known with great accuracy. This is especially true in data science where statistical generalization error leads to very noisy observations of the ground truth. It also occurs in science and engineering, like vibrational studies, where the data is subject to modeling and measurements errors. All this uncertainty can have a large impact on the nominal solution, that is, the optimal solution in case the data is treated as exact, and thereby making that solution less useful from a practical point of view. This is well known in the field of robust optimization; see [5] for specific examples in linear programming. In data science, it causes overfitting which reduces the generalization of the optimized problem to unobserved data. There exists therefore a need to obtain *robust solutions* that are more immune to such uncertainty.

In this paper, we propose to obtain robust solutions to (1) by optimizing for the worst-case behavior. This is a popular paradigm for convex optimization problems (see the book [4] for an overview) but to the best of our knowledge a systematic treatment is new for the RQ. To represent uncertainties in (1), we let the entries of A and B depend on some parameters $\mu \in \mathbb{R}^m$ and $\xi \in \mathbb{R}^p$. Specifically, we consider A and B as the smooth matrix-valued functions

$$\begin{aligned} A: &\quad \mu \in \Omega \ \mapsto \ A(\mu) \in \mathbb{R}^{n\times n}, \\ B: &\quad \xi \in \Gamma \ \mapsto \ B(\xi) \in \mathbb{R}^{n\times n}, \end{aligned} \tag{2}$$

where $A(\mu) \succ 0$ and $B(\xi) \succeq 0$ for all $\mu \in \Omega$ and $\xi \in \Gamma$, and $\Omega \subset \mathbb{R}^m$ and $\Gamma \subset \mathbb{R}^p$ are compact. To take into account the uncertainties, we minimize the worst-case RQ

$$\min_{\substack{z\in\mathbb{R}^n \\ z\neq 0}} \max_{\substack{\mu\in\Omega \\ \xi\in\Gamma}} \frac{z^T A(\mu) z}{z^T B(\xi) z}. \tag{3}$$

We call (3) a *robust RQ minimization* problem. Since $B(\xi)$ is only positive semidefinite, the inner max problem might yield a value $+\infty$ for particular z. For well-posedness, we assume $B(\xi) \not\equiv 0$, so a non-trivial finite minimizer of (3) exists. The positive definiteness constraint of A will be exploited in this paper, but they can be relaxed; see Remark 2 later.

By denoting the optimizers[1] of the max problem as

$$\mu^*(z) := \arg\max_{\mu\in\Omega} z^T A(\mu) z \quad \text{and} \quad \xi^*(z) := \arg\min_{\xi\in\Gamma} z^T B(\xi) z, \tag{4}$$

and introducing the coefficient matrices

$$G(z) := A(\mu^*(z)) \quad \text{and} \quad H(z) := B(\xi^*(z)), \tag{5}$$

we can write the problem (3) as

$$\min_{\substack{z\in\mathbb{R}^n \\ z\neq 0}} \frac{\max_{\mu\in\Omega} z^T A(\mu) z}{\min_{\xi\in\Gamma} z^T B(\xi) z} = \min_{\substack{z\in\mathbb{R}^n \\ z\neq 0}} \frac{z^T G(z) z}{z^T H(z) z}. \tag{6}$$

This is a nonlinear RQ minimization problem with coefficient matrices depending nonlinearly on the vector z.

[1] When the optimizers are not unique, μ^* and ξ^* denote any of them.

1.2 Background and applications

The robust RQ minimization occurs in a number of applications. For example, it is of particular interest in *robust adaptive beamforming* for array signal processing in wireless communications, medical imaging, radar, sonar and seismology, see, e.g., [17]. A standard technique in this field to make the optimized beamformer less sensitive to errors caused by imprecise sensor calibrations is to explicitly use uncertainty sets during the optimization process. In some cases, this leads to robust RQ optimization problems, see [30, 24] for explicit examples that are of the form (3). For simple uncertainty sets, the robust RQ problem can be solved in closed-form. This is however no longer true for more general uncertainty sets, showing the necessity of the algorithms proposed in this paper.

Fisher's linear discriminant analysis (LDA) was extended in [15] to allow for general convex uncertainty models on the data. The resulting *robust LDA* is the corresponding worst-case optimization problem for LDA and is an example of (3). For product type uncertainty with ellipsoidal constraints, the inner maximization can be solved explicitly and thus leads to a problem of the form (6). Since the objective function is of fractional form, the robust Fisher LDA can be solved by convex optimization as in [14], where the same technique is also used for robust matched filtering and robust portfolio selection [14]. As we will show in numerical experiments, it might be beneficial to solve the robust LDA by other algorithms than those that explicitly exploit convexity. In addition, convexity is rarely present in more realistic problems.

Generalized eigenvalue classifier (GEC) determines two hyperplanes to distinguish two classes of data [20]. When the data points are subject to ellipsoidal uncertainty, the resulting worst-case analysis problem can again be written as a robust RQ problem of the form (6). This formulation is used explicitly in [31] for the solution of the *robust GEC*.

Common spatial pattern (CSP) analysis is routinely applied in feature extraction of electroencephalogram data in brain-computer interface systems, see, e.g., [7]. The *robust common spatial filters* studied in [13] is another example of (6) where the uncertainty on the covariance matrices is of product type.

Robust GEC and CSP cannot be solved by convex optimization. Fortunately, since (6) is a nonlinear RQ problem, it is a natural idea to solve it by the following fixed-point iteration scheme:

$$z_{k+1} \longleftarrow \operatorname*{arg\,min}_{\substack{z\in\mathbb{R}^n \\ z\neq 0}} \frac{z^T G(z_k) z}{z^T H(z_k) z}, \quad k = 0, 1, \ldots \tag{7}$$

In each iteration, a standard RQ minimization problem, that is, an eigenvalue problem, is solved. This simple iterative scheme is widely used in other fields as well. In computational physics and chemistry it is known as the self-consistent-field (SCF) iteration (see, e.g., [16]). Its convergence behavior applied to (6) remains however unknown.

A block version of the robust RQ minimization can be found in [2]. A similar problem with a finite uncertainty set is considered in [27, 25]. A different treatment of uncertainty in RQ minimization occurs in uncertainty quantification. Contrary to our minimax approach, the aim there is to compute statistical properties (e.g., moments) of the solution of a stochastic eigenvalue problem given the probability distribution of the data matrices. While the availability of the random solution is appealing, it also leads to a much more computationally demanding problem; see [9, 6] for recent development of stochastic eigenvalue problems. Robust RQ problems on the other hand can be solved at the expense of typically only a few eigenvalue computations.

1.3 Contributions and outline

In this paper, we propose to solve the robust RQ minimization problem using techniques from nonlinear eigenvalue problems. We show that the nonlinear RQ minimization problem (6) can be characterized as a nonlinear eigenvalue problem with eigenvector dependence (NEPv). We explain that the simple iterative scheme (7) can fail to converge due to a wrong ordering of the eigenvalues, and we will show how to solve this issue by a nonlinear spectral transformation. By taking into account the second-order derivatives of the nonlinear RQ, we derive a modified NEPv and prove that the simple iterative scheme (7) for the modified NEPv is locally quadratic convergent. Finally, we discuss applications and detailed numerical examples in data science applied to a variety of data sets. The numerical experiments clearly show that a robust solution can be computed efficiently with the proposed methods. In addition, our proposed algorithms typically generate better optimizers measured by cross-validation than the ones from the traditional methods, like simple fixed-point iteration.

The paper is organized as follows. In §2, we study basic properties of the coefficient matrices $G(z)$ and $H(z)$. In §3, we derive NEPv characterizations of the nonlinear RQ optimization problem (6). In §5, we discuss three applications of the robust RQ minimization. Numerical examples for these applications are presented in §6. Concluding remarks are in §7

Notations: Throughout the paper, we follow the notation commonly used in numerical linear algebra. We call λ an eigenvalue of a matrix pair (A, B) with an associated eigenvector x, if both satisfy the generalized linear eigenvalue problem $Ax = \lambda Bx$. We call $\lambda = \infty$ an eigenvalue if $Bx = \lambda^{-1}Ax = 0$. An eigenvalue λ is called simple, if its algebraic multiplicity is one. When A and B are symmetric with either $A \succ 0$ or $B \succ 0$, we call (A, B) a symmetric definite pair and we use $\lambda_{\min}(A, B)$ and $\lambda_{\max}(A, B)$ to denote the minimum and maximum of its (real) eigenvalues, respectively.

2 Basic properties

In the following, we first consider basic properties of the coefficient matrices $G(z)$ and $H(z)$ defined in (5), as well as the numerator and denominator of the nonlinear RQ (6).

Lemma 1. *(a) For a fixed $z \in \mathbb{R}^n$, $G(z) = G(z)^T \succ 0$ and $H(z) = H(z)^T \succeq 0$.*

(b) $G(z)$ and $H(z)$ are homogeneous matrix functions in $z \in \mathbb{R}^n$, i.e., $G(\alpha z) = G(z)$ and $H(\alpha z) = H(z)$ for $\alpha \neq 0$ and $\alpha \in \mathbb{R}$.

(c) The numerator $g(z) = z^T G(z)z$ of (6) is a strongly convex function in z. In particular, if $g(z)$ is smooth at z, then $\nabla^2 g(z) \succ 0$.

Proof. (a) Follows from $A(\mu) \succ 0$ for $\mu \in \Omega$. Hence, $G(z) = A(\mu^*(z))$ is also symmetric positive definite. We can show $H(z) \succeq 0$ in analogy.

(b) Follows from $\max_{\mu\in\Omega}(\alpha z)^T A(\mu)(\alpha z) = \alpha^2 \max_{\mu\in\Omega} z^T A(\mu)z$ which implies $\mu^*(\alpha z) = \mu^*(z)$ for all $\alpha \neq 0$. Hence, $G(\alpha z) = G(z)$ is homogeneous in z. We can show $H(\alpha z) = H(z)$ in analogy.

(c) Since $A(\mu) \succ 0$ for $\mu \in \Omega$ and Ω is compact, the function $g_\mu(z) = z^T A(\mu)z$ is a strongly convex function in z with uniformly bounded $\lambda_{\min}(\nabla^2 g_\mu(z)) = 2 \cdot \lambda_{\min}(A(\mu)) \geq \delta > 0$ for all $\mu \in \Omega$. Hence, the pointwise maximum $g(z) = \max\limits_{\mu\in\Omega} g_\mu(z)$ is strongly convex as well. □

A situation of particular interest is when z satisfies the following regularity condition.

Definition 1 (Regularity). *A point $z \in \mathbb{R}^n$ is called* regular *if $z \neq 0$, $z^T H(z) z \neq 0$, and the functions $\mu^*(z)$ and $\xi^*(z)$ in (4) are twice continuously differentiable at z.*

Regularity is not guaranteed from the formulation of minimax problem (3). However, we observe that it is not a severe restriction in applications where the optimal parameters $\mu^*(z)$ and $\xi^*(z)$ have explicit and analytic expressions; see §6.

When z is regular, both $G(z)$ and $H(z)$ are smooth matrix-valued functions at z. It allows us to define the gradient of numerator and denominator functions

$$g(z) = z^T G(z) z \quad \text{and} \quad h(z) = z^T H(z) z$$

of the nonlinear RQ (6). By straightforward calculations and using the symmetry of $G(z)$ and $H(z)$, we obtain

$$\nabla g(z) = (2G(z) + \widetilde{G}(z))z \quad \text{and} \quad \nabla h(z) = (2H(z) + \widetilde{H}(z))z, \tag{8}$$

where

$$\widetilde{G}(z) = \begin{bmatrix} z^T \frac{\partial G(z)}{\partial z_1} \\ z^T \frac{\partial G(z)}{\partial z_2} \\ \vdots \\ z^T \frac{\partial G(z)}{\partial z_n} \end{bmatrix} \quad \text{and} \quad \widetilde{H}(z) = \begin{bmatrix} z^T \frac{\partial H(z)}{\partial z_1} \\ z^T \frac{\partial H(z)}{\partial z_2} \\ \vdots \\ z^T \frac{\partial H(z)}{\partial z_n} \end{bmatrix}. \tag{9}$$

Lemma 2. *Let $z \in \mathbb{R}^n$ be regular. The following results hold.*

(a) $z^T \frac{\partial G}{\partial z_i}(z) z \equiv 0$ and $z^T \frac{\partial H}{\partial z_i}(z) z \equiv 0$, for $i = 1, \dots, n$.

(b) $\nabla g(z) = 2G(z)z$ and $\nabla^2 g(z) = 2\left(G(z) + \widetilde{G}(z)\right)$.

(c) $\nabla h(z) = 2H(z)z$ and $\nabla^2 h(z) = 2\left(H(z) + \widetilde{H}(z)\right)$.

Proof. (a) By definition of $G(z)$ and smoothness of $\mu^*(z)$ and $A(\mu)$, we have

$$z^T \frac{\partial G}{\partial z_i}(z) z = z^T \frac{\partial A(\mu^*(z))}{\partial z_i} z = z^T \left(\frac{dA(\mu^*(z + te_i))}{dt} \bigg|_{t=0} \right) z,$$

where e_i is the ith column of the $n \times n$ identity matrix. Introducing $f(t) = z^T A(\mu^*(z + te_i))z$, we have $z^T \frac{\partial G}{\partial z_i}(z) z = f'(0)$. Since

$$f(0) = z^T A(\mu^*(z))z = \max_{\mu \in \Omega} z^T A(\mu) z \geq z^T A(\mu^*(z + te_i)\mu)z = f(t),$$

the smooth function $f(t)$ achieves its maximum at $t = 0$. Hence, $f'(0) = 0$ and the result follows. The proof for $H(z)$ is completely analogous.

(b) The result $\nabla g(z) = 2G(z)z$ follows from equation (8) and result (a). Continuing from this equation, we obtain $\nabla^2 g(z) = 2G(z) + 2\widetilde{G}(z)^T$. Since the Hessian $\nabla^2 g(z)$ is symmetric, we have that $\widetilde{G}(z)$ is a symmetric matrix due to the symmetry of $G(z)$.

(c) The proof is similar to that of (b). □

We can see that the gradients of $g(z)$ and $h(z)$ in (8) are simplified due to the null vector property by Lemma 2(a):

$$\widetilde{G}(z)z = 0 \quad \text{and} \quad \widetilde{H}(z)z = 0 \quad \text{for all regular } z \in \mathbb{R}^n. \tag{10}$$

In addition, from the proof of Lemma 2, we can also see that both $\widetilde{G}(z)$ and $\widetilde{H}(z)$ are symmetric matrices. The symmetry property is not directly apparent from definition (9), but it is implied by the optimality of μ^* and ξ^* in (4). This property will be useful in the discussion of the nonlinear eigenvalue problems in the following section.

3 Nonlinear eigenvalue problems

In this section, we characterize the nonlinear RQ minimization problem (6) as two different nonlinear eigenvalue problems. First, we note that the homogeneity of $G(z)$ and $H(z)$ from Lemma 1(b) and the positive semidefiniteness of $H(z)$ allow us to rewrite (6) as the constrained minimization problem

$$\min_{z\in\mathbb{R}^n} z^T G(z) z \quad \text{s.t.} \quad z^T H(z) z = 1. \tag{11}$$

The characterization will then follow from the stationary conditions of this constrained problem. For this purpose, let us define its Lagrangian function with multiplier λ:

$$L(z,\lambda) = z^T G(z) z - \lambda(z^T H(z) z - 1).$$

Theorem 1 (First-order NEPv)**.** *Let $z \in \mathbb{R}^n$ be regular. A necessary condition for z being a local optimizer of (11) is that z is an eigenvector of the nonlinear eigenvalue problem*

$$G(z)z = \lambda H(z) z, \tag{12}$$

for some (scalar) eigenvalue λ.

Proof. This result follows directly from the first-order optimality conditions [22, Theorem 12.1] of the constrained minimization problem (11),

$$\nabla_z L(z,\lambda) = 0 \quad \text{and} \quad z^T H(z) z = 1, \tag{13}$$

combined with the gradient formulas (b) and (c) in Lemma 2. □

Any (λ, z) with $z \neq 0$ and $\lambda < \infty$ satisfying equation (12) is called an eigenpair of the NEPv. This means that the vector z must also be an eigenvector of the matrix pair $(G(z), H(z))$. Since $G(z) \succ 0$ and $H(z) \succeq 0$ are symmetric, the matrix pair has n linearly independent eigenvectors and n strictly positive (counting ∞) eigenvalues. However, Theorem 1 does not specify to which eigenvalue that z corresponds. To resolve this issue, let us take into account the second-order derivative information.

Theorem 2 (Second-order NEPv)**.** *Let $z \in \mathbb{R}^n$ be regular and define $\mathcal{G}(z) = G(z) + \widetilde{G}(z)$ and $\mathcal{H}(z) = H(z) + \widetilde{H}(z)$. A necessary condition for z being a local minimizer of (11) is that it is an eigenvector of the nonlinear eigenvalue problem*

$$\mathcal{G}(z) z = \lambda \mathcal{H}(z) z, \tag{14}$$

corresponding to the smallest positive eigenvalue λ of the matrix pair $(\mathcal{G}(z), \mathcal{H}(z))$ with $\mathcal{G}(z) \succ 0$. Moreover, if λ is simple, then this condition is also sufficient.

Proof. By the null vector properties in (10), we see that the first- and second-order NEPvs (12) and (14) share the same eigenvalue λ and eigenvector z,

$$\mathcal{G}(z)z - \lambda\mathcal{H}(z)z = G(z)z - \lambda H(z) z = 0.$$

Hence, by Theorem 1 if z is a local minimizer of (11), it is also an eigenvector of (14).

It remains to show the order of the corresponding eigenvalue λ. Both $\mathcal{G}(z)$ and $\mathcal{H}(z)$ are symmetric by Lemma 2, and $\mathcal{G}(z) \succ 0$ is also positive definite by Lemma 1(c). Hence, the eigenvalues of the pair $(\mathcal{G}(z), \mathcal{H}(z))$ are real (including infinity eigenvalues) and we can denote them as

$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Let $v^{(1)}, v^{(2)}, \dots, v^{(n)}$ be their corresponding $\mathcal{G}(z)$-orthogonal[2] eigenvectors. Since (λ, z) is an eigenpair of both (12) and (14), we have for some finite eigenvalue λ_j that

$$z = v^{(j)}, \quad \lambda = \lambda_j > 0.$$

We will deduce the order of λ_j from the second-order necessary condition of (11) for the local minimizer z (see, e.g., [22, Theorem 12.5]):

$$s^T \nabla_{zz} L(z, \lambda) s \geq 0, \quad \forall s \in \mathbb{R}^n \text{ s.t. } s^T H(z) z = 0. \tag{15}$$

Take any $s \in \mathbb{R}^n$ such that $s^T H(z) z = 0$, its expansion in the basis of eigenvectors satisfies

$$s = \sum_{i=1}^{n} \alpha_i v^{(i)} = \sum_{i=1, i \neq j}^{n} \alpha_i v^{(i)},$$

since $\alpha_j = s^T \mathcal{G}(z) z = \lambda_j s^T \mathcal{H}(z) z = \lambda_j s^T H(z) z = 0$. Using the Hessian formulas (b) and (c) in Lemma 2, the inequality in (15) can be written as

$$s^T \left(\nabla^2 g(z) - \lambda \nabla^2 h(z) \right) s = 2 \cdot s^T \left(\mathcal{G}(z) - \lambda \mathcal{H}(z) \right) s \geq 0.$$

Combining with the expansion of s and $\lambda = \lambda_j$, the inequality above becomes

$$\sum_{i=1, i \neq j}^{n} \alpha_i^2 \left(1 - \frac{\lambda_j}{\lambda_i} \right) \geq 0, \tag{16}$$

which holds for all $\alpha_1, \dots, \alpha_n$. The necessary condition (15) is therefore equivalent to

$$1 - \frac{\lambda_j}{\lambda_i} \geq 0 \quad \forall i = 1, \dots, n \text{ with } i \neq j.$$

Since $\lambda = \lambda_j > 0$, we have shown that $0 < \lambda \leq \lambda_i$ for all positive eigenvalues $\lambda_i > 0$.

Finally, if the smallest positive eigenvalue $\lambda = \lambda_j$ is simple, then for any $s \in \mathbb{R}^n$ such that $s \neq 0$ and $s^T H(z) z = 0$, the inequalities from above lead to

$$\sum_{i=1, i \neq j}^{n} \alpha_i^2 \left(1 - \frac{\lambda_j}{\lambda_i} \right) = s^T \nabla_{zz} L(z, \lambda) s > 0. \tag{17}$$

We complete the proof by noticing that (17) corresponds to the second-order sufficient condition for z being a strict local minimizer of (11)

$$s^T \nabla_{zz} L(z, \lambda) s \ > 0, \quad \forall s \in \mathbb{R}^n \text{ s.t. } s^T H(z) z = 0 \text{ and } s \neq 0,$$

see, e.g., [22, Theorem 12.6]. □

By Theorem 1 (first-order NEPv), we see that a regular local minimizer z of the nonlinear RQ (6) is an eigenvector of the matrix pair $(G(z), H(z))$. Although the pair has positive eigenvalues, we do not know to which one z belongs. On the other hand, by Theorem 2 (second-order NEPv), the same vector z is also an eigenvector of the matrix pair $(\mathcal{G}(z), \mathcal{H}(z))$. This pair has real eigenvalues that are not necessarily positive, but we know that z belongs to the smallest strictly positive eigenvalue. Simplicity of this eigenvalue also guarantees that z is a strict local minimizer of (6).

[2] $(v^{(i)})^T \mathcal{G}(z) v^{(j)} = \delta_{ij}$ for $i, j = 1, \dots, n$, where $\delta_{ij} = 0$ if $i \neq j$, otherwise $\delta_{ij} = 1$.

Remark 1. Since $\mathcal{G}(z)$ is symmetric positive definite but $\mathcal{H}(z)$ is only symmetric, it is numerically advisable to compute the eigenvalues λ of the pair $(\mathcal{G}(z), \mathcal{H}(z))$ as the eigenvalues λ^{-1} of the symmetric definite pair $(\mathcal{H}(z), \mathcal{G}(z))$.

Remark 2. From the proofs of Theorems 1 and 2, we can see that it is also possible to derive NEPv characterizations if we relax the positive definite conditions (2) to $A(\mu) \succeq 0$, $B(\xi) \succeq 0$ and $A(\mu) + B(\xi) \succ 0$ for $\mu \in \Omega$ and $\xi \in \Gamma$. The last condition is to guarantee the well-posedness of the ratio $\frac{z^T A(\mu) z}{z^T B(\xi) z}$ by avoiding the case $0/0$, i.e., $z^T A(\mu) z = z^T B(\xi) z = 0$. For Theorem 2 to hold, we need to further assume that $A(\mu^*(z)) \succ 0$. This guarantees $\mathcal{G}(z) \succ 0$ and the eigenvalue $\lambda > 0$.

4 SCF iterations

The coefficient matrices in the eigenvalue problems (12) and (14) depend nonlinearly on the eigenvector, hence they are nonlinear eigenvalue problems with eigenvector dependence. Such type of problems also arise in the Kohn–Sham density functional theory in electronic structure calculations [21], the Gross–Pitaevskii equation for modeling particles in the state of matter called the Bose-Einstein condensate [12], and linear discriminant analysis in machine learning [36]. As mentioned in the introduction, a popular algorithm to solve such NEPv is the SCF iteration. The basic idea is that, by fixing the eigenvector dependence in the coefficient matrices, we end up with a standard eigenvalue problem. Iterating on the eigenvector then gives rise to SCF.

Applied to the NEPv (12) or (14), the SCF iteration is

$$z_{k+1} \longleftarrow \text{an eigenvector of } G_k z = \lambda\, H_k z, \tag{18}$$

where $G_k = G(z_k)$ and $H_k = H(z_k)$ for the first-order NEPv (12), or $G_k = \mathcal{G}(z_k)$ and $H_k = \mathcal{H}(z_k)$ for the second-order NEPv (14), respectively. For the second-order NEPv, it is clear by Theorem 2 that the update z_{k+1} should be the eigenvector corresponding to the smallest positive eigenvalue of the matrix pair $(\mathcal{G}(z_k), \mathcal{H}(z_k))$. However, for the first-order NEPv, we need to decide which eigenvector of the matrix pair $(G_k,\, H_k)$ to use for the update z_{k+1}. We will address this so-called *eigenvalue ordering issue* in next subsection.

4.1 A spectral transformation for the first-order NEPv

Since the first-order NEPv (12) is related to the optimality conditions of the minimization problem (11), it seems natural to choose z_{k+1} in (18) as the eigenvector belonging to the smallest eigenvalue of the matrix pair $(G(z_k), H(z_k))$. After all, our main interest is minimizing the robust RQ in (3), for which the simple fixed-point scheme (7) indeed directly leads to the SCF iteration

$$z_{k+1} \longleftarrow \text{ eigenvector of the smallest eigenvalue of } G(z_k)z = \lambda\, H(z_k)z. \tag{19}$$

In order to have any hope for convergence of (19) as $k \to \infty$, there needs to exist an eigenpair (λ_*, z_*) of the NEPv that corresponds to the smallest eigenvalue λ_* of the matrix pair $(G(z_*),\, H(z_*))$, that is,

$$G(z_*)z_* = \lambda_* H(z_*) z_* \quad \text{with} \quad \lambda_* = \lambda_{\min}(G(z_*),\, H(z_*)).$$

(Indeed, simply take $z_0 = z_*$). Unfortunately, there is little theoretical justification for this in the case of the robust RQ minimization. As we will show in the numerical experiments in §6, it fails to hold in practice.

In order to deal with this eigenvalue ordering issue, we propose the following (nonlinear) spectral transformation:

$$G_\sigma(z)z = \mu H(z)z, \tag{20}$$

where

$$G_\sigma(z) = G(z) - \sigma(z) \cdot \frac{H(z)zz^T H(z)}{z^T H(z)z},$$

and $\sigma(z)$ is a scalar function in z. It is easy to see that the nonlinearly shifted NEPv (20) is equivalent to the original NEPv (12) in the sense that $G(z)z = \lambda H(z)z$ if and only if

$$G_\sigma(z)z = \mu H(z)z \quad \text{with} \quad \mu = \lambda - \sigma(z). \tag{21}$$

The lemma below shows that with a proper choice of the shift function $\sigma(z)$, the eigenvalue μ of the shifted first-order NEPv (20) will be the smallest eigenvalue of the matrix pair $(G_\sigma(z), H(z))$.

Lemma 3. *Let $\beta > 1$ and define the shift function*

$$\sigma(z) = \beta \cdot \lambda_{\max}(G(z),\, H(z)) - \lambda_{\min}(G(z),\, H(z)). \tag{22}$$

If $(\mu,\, z)$ is an eigenpair of the shifted first-order NEPv (20), then μ is the simple smallest eigenvalue of the matrix pair $(G_\sigma(z),\, H(z))$.

Proof. Similar to the proof of Theorem 2, let $0 < \lambda_1 \leq \cdots \leq \lambda_n$ be the n eigenvalues of the pair $(G(z), H(z))$ with corresponding $G(z)$-orthogonal eigenvectors $v^{(1)}, \ldots, v^{(n)}$. By (21), we know that if $(\mu,\, z)$ is an eigenpair of (20), it implies that $\lambda_j = \mu + \sigma(z)$ and $z = v^{(j)}$ for some j.

From $G(z)v^{(i)} = \lambda_i H(z)v^{(i)}$, we obtain $z^T H(z)v^{(i)} = \delta_{ij}\lambda_i^{-1}$. Hence, using $z = v^{(j)}$ it holds

$$G_\sigma(z)v^{(i)} = G(z)v^{(i)} = \lambda_i H(z)v^{(i)} \quad \text{for} \quad i \neq j,$$

and

$$G_\sigma(z)v^{(j)} = G(z)v^{(j)} - \sigma(z) \cdot H(z)v^{(j)} = (\lambda_j - \sigma(z)) \cdot H(z)v^{(j)}.$$

So the n eigenvalues of the shifted pair $(G_\sigma(z), H(z))$ are given by λ_i for $i \neq j$ and $\mu = \lambda_j - \sigma(z)$. By construction of $\sigma(z)$ in (22), we also have

$$\mu = \lambda_j - \sigma(z) = \lambda_j - (\beta\lambda_n - \lambda_1) < \lambda_j - (\lambda_n - \lambda_i) \leq \lambda_i \quad \text{for all } i \neq j.$$

So μ is indeed the simple smallest eigenvalue of the shifted pair, and we complete the proof. □

A similar idea to the nonlinear spectral transformation (20) is the level shifting [23] and the trust-region SCF [29, 33] schemes to solve NEPv from quantum chemistry. However, the purpose of these schemes is to stabilize the SCF iteration and not to reorder the eigenvalues as we do here. Lemma 3 provides a justification to apply SCF iteration to the shifted first-order NEPv (20) and take z_{k+1} as the eigenvector corresponding to the smallest eigenvalue. This procedure is summarized in Alg. 1. The optional line search in line 6 will be discussed in §4.3.

In Alg. 1 we have chosen $\beta = 1.01$ for convenience. In practice there seems to be little reason for choosing larger values of β. This can be intuitively understood from the fact that we can rewrite the solution z_{k+1} in Alg. 1 as follows:

$$z_{k+1} \longleftarrow \arg \min_{z^T H(z_k) z = 1} \left\{ \frac{z^T G(z_k) z}{z^T H(z_k) z} + \frac{\sigma_k}{2} \left\| H^{1/2}(z_k) \left(zz^T - z_k z_k^T \right) H^{1/2}(z_k) \right\|_F^2 \right\}, \tag{23}$$

Algorithm 1 SCF iteration for the shifted first-order NEPv (20).

```
Input: initial $z_0 \in \mathbb{R}^n$, tolerance tol, shift factor $\beta > 1$ (e.g. $\beta = 1.01$).
Output: approximate eigenvector $\widehat{z}$.
for $k = 0, 1, \ldots$ do
    set $G_k = G(z_k)$, $H_k = H(z_k)$, and $\rho_k = z_k^T G_k z_k / (z_k^T H_k z_k)$.
    if $\|G_k z_k - \rho_k H_k z_k\|_2 / (\|G_k z_k\|_2 + \rho_k \|H_k z_k\|_2) \leq$ tol then return $\widehat{z} = z_{k+1}$.
    shift $G_{\sigma k} = G_k - \frac{\sigma_k}{z_k^T H_k z_k} H_k z_k z_k^T H_k$ with $\sigma_k = \beta \cdot \lambda_{\max}(G_k, H_k) - \lambda_{\min}(G_k, H_k)$.
    compute the smallest eigenvalue and eigenvector $(\mu_{k+1}, z_{k+1})$ of $(G_{\sigma k}, H_k)$.
    (optional) perform line search to obtain $z_{k+1}$.
end for
```

where z_k is assumed to be normalized as $z_k^T H(z_k) z_k = 1$. Observe that the σ_k-term in (23) is the distance of z to z_k expressed in a weighted norm since $H(z_k) \succ 0$, and z and z_k are normalized vectors. Hence, in each iteration of Alg. 1 we solve a penalized RQ minimization where the penalization promotes that z_{k+1} is close to z_k. Therefore, we may encounter slower convergence with a larger penalization factor σ_k and thus also a larger β. From the viewpoint of solving penalized RQ minimization (23), a smaller penalty σ_k (and therefore a smaller β) on the step size can also lead to a faster convergence compared to what we will observe in the numerical experiments from §6. However, it is not theoretically guaranteed.

4.2 Local convergence of SCF iteration for the second-order NEPv

Thanks to Theorem 2, we can apply SCF iteration directly to (14) while targeting the smallest positive eigenvalue in each iteration. This procedure is summarized in Alg. 2.

Algorithm 2 SCF iteration for the second-order NEPv (14).

```
Input: initial $z_0 \in \mathbb{R}^n$, tolerance tol.
Output: approximate eigenvector $\widehat{z}$.
1: for $k = 0, 1, \ldots$ do
2:     set $G_k = \mathcal{G}(z_k)$, $H_k = \mathcal{H}(z_k)$ and $\rho_k = z_k^T G_k z_k / (z_k^T H_k z_k)$.
3:     if $\|G_k z_k - \rho_k H_k z_k\|_2 / (\|G_k z_k\|_2 + \rho_k \|H_k z_k\|_2) \leq$ tol then return $\widehat{z} = z_{k+1}$.
4:     compute the smallest strictly positive eigenvalue and eigenvector $(\lambda_{k+1}, z_{k+1})$ of $(G_k, H_k)$.
5:     (optional) perform line search to obtain $z_{k+1}$.
6: end for
```

Due to the use of second-order derivatives, one would hope that the local convergence rate is at least quadratic. The next theorem shows exactly that.

Theorem 3 (Quadratic convergence). *Let (λ, z) be an eigenpair of the second-order NEPv* (14) *such that λ is simple and the smallest eigenvalue of the matrix pair $(\mathcal{G}(z), \mathcal{H}(z))$. If z_k in Alg.* 2 *is such that $|\sin\angle(z_k, z)|$ is sufficiently small, then the iterate z_{k+1} in line 4 satisfies*

$$\sin\angle(z_{k+1}, z) = O(|\sin\angle(z_k, z)|^2),$$

where $\angle(u, v)$ is the angle between the vectors u and v.

Proof. For clarity, let us denote the eigenpair of the second-order NEPv (14) as (λ_*, z_*). Due to the homogeneity of $\mathcal{G}(z)$ and $\mathcal{H}(z)$ in z, we can always assume that $\|z_*\|_2 = \|z_k\|_2 = \|z_{k+1}\|_2 = 1$. Hence,

$$z_k = z_* + d \quad \text{with} \quad \|d\|_2 = 2\sin\left(\tfrac{1}{2}\angle(z_k, z_*)\right) = O(|\sin\angle(z_k, z_*)|).$$

We will regard the eigenpair (λ_{k+1}, z_{k+1}) of $(\mathcal{G}(z_k), \mathcal{H}(z_k))$ as a perturbation of the eigenpair (λ_*, z_*) of $(\mathcal{G}(z_*), \mathcal{H}(z_*))$. This is possible since the matrix $\mathcal{G}(z_*)$ is positive definite and, for $\|d\|_2$ sufficiently small, λ_{k+1} remains simple and it will be the closest eigenvalue of λ_*. Denoting

$$g_k = (\mathcal{G}(z_k) - \mathcal{G}(z_*))\, z_*, \quad h_k = (\mathcal{H}(z_k) - \mathcal{H}(z_*))\, z_*,$$

we apply the standard eigenvector perturbation analysis for definite pairs; see, e.g. [28, Theorem VI.3.7]. This gives together with the continuity of $\mathcal{G}(z)$ and $\mathcal{H}(z)$ that

$$\|z_{k+1} - \alpha z_*\|_2 \leq c \cdot \max(\|g_k\|_2, \|h_k\|_2),$$

where $|\alpha| = 1$ is a rotation factor and c is a constant depending only on the gap between λ_* and the rest of the eigenvalues of $(\mathcal{G}(z_*), \mathcal{H}(z_*))$. To complete the proof it remains to show that

$$\|g_k\|_2 = O(\|d\|_2^2) \quad \text{and} \quad \|h_k\|_2 = O(\|d\|_2^2).$$

The result for g_k follows from $\mathcal{G}(z_*)z_* = G(z_*)z_*$ and the Taylor expansion

$$G(z_*)z_* = G(z_k)z_k - \mathcal{G}(z_k)d + O(\|d\|_2^2) = \mathcal{G}(z_k)z_* + O(\|d\|_2^2),$$

where we have used $\nabla(G(z)z) = \frac{1}{2}\nabla^2 g(z) = \mathcal{G}(z)$ from Lemma 2(b). The result for h_k is derived analogously. □

The convergence of SCF iteration has been studied for NEPv arising in electronic structure calculations and machine learning, where local linear convergence is proved under certain assumptions; see, e.g., [16, 32, 19, 35]. Our first-order and second-order NEPv, however, do not fall in this category and hence these analyses do not carry over directly. However, thanks to the special structure of the second-order NEPv, we can for instance prove the local quadratic convergence. Our numerical experience suggests that Alg. 2 usually converges much faster than the first-order Alg. 1. But it also requires the derivatives of $G(z)$ and $H(z)$ in order to generate the coefficient matrices $\mathcal{G}(z)$ and $\mathcal{H}(z)$. In cases where those matrices are not available or too expensive to compute, one can still resort to Alg. 1 for the solution.

4.3 Implementation issues

The SCF iteration is not guaranteed to be monotonic in the nonlinear RQ

$$\rho(z) = \frac{z^T G(z) z}{z^T H(z) z}.$$

This is a common issue for SCF iterations. A simple remedy is to apply damping for the update z_{k+1}:

$$z_{k+1} \longleftarrow \alpha z_{k+1} + (1-\alpha) z_k, \tag{24}$$

where $0 \leq \alpha \leq 1$ is a damping factor. This is also called the mixing scheme when applied to the density matrix $z_k z_k^T$ instead of to z_k in electronic structure calculations [16]. Ideally, one would like to choose α so that it leads to the optimal value of the nonlinear RQ

$$\min_{\alpha\in[0,1]} \rho(\alpha z_{k+1} + (1-\alpha) z_k).$$

Since an explicit formula for the optimal α is usually unavailable, one has to instead apply a line search for $\alpha \in [0, 1]$ to obtain

$$\rho(z_k + \alpha d_k) < \rho(z_k) \quad \text{with} \quad d_k = z_{k+1} - z_k. \tag{25}$$

See, for example, Alg. 3 where this is done by Armijo backtracking. Such a line search works if d_k is a descent direction of $\rho(z)$ at z_k, i.e.,

$$d_k^T \nabla\rho(z_k) < 0 \quad \text{with} \quad \nabla\rho(z) = \frac{2}{z^T H(z) z}\Big(G(z) - \rho(z)H(z)\Big)z.$$

As long as $z_{k+1}^T H_k z_k \neq 0$, we can always obtain such a direction by suitably (scalar) normalizing the eigenvectors z_{k+1} to satisfy

(a) Line 5 of Alg. 1: $z_{k+1}^T H_k z_k > 0$, which leads to

$$\tfrac{1}{2} d_k^T \nabla\rho(z_k) = \Big(\mu_{k+1} + \sigma_k - \rho_k\Big) \cdot \zeta_k < 0,$$

(b) Line 4 of Alg. 2: $(\lambda_{k+1} - \rho_k) \cdot z_{k+1}^T H_k z_k > 0$, which leads to

$$\tfrac{1}{2} d_k^T \nabla\rho(z_k) = (\lambda_{k+1} - \rho_k) \cdot \zeta_k < 0.$$

Here, $\zeta_k = (z_{k+1}^T H_k z_k)/(z_k^T H_k z_k)$, and for (a) we exploited that z_k is not an eigenvector of (G_k, H_k) (in which case, Alg. 1 stops at line 3), so that $\mu_{k+1} = \min_z \frac{z^T G_{\sigma k} z}{z^T H_k z} < \frac{z_k^T G_{\sigma k} z_k}{z_k^T H_k z_k} = -\sigma_k + \rho_k$.

Algorithm 3 Line search

```
Input: starting point $z_k$, descent direction $d_k$, factors $c, \tau \in (0,1)$ (e.g., $c = \tau = 0.1$).
Output: $z_{k+1} = z_k + \alpha d_k$.
1: Set $\alpha = 1$ and $t = -cm$ with $m = d_k^T \nabla\rho(z_k)$.
2: while $\rho(z_k) - \rho(z_k + \alpha d_k) < \alpha t$ do
3:     $\alpha := \tau\alpha$
4: end while
```

In the rare case of $z_{k+1}^T H_k z_k = 0$, the increment $d_k = z_{k+1} - z_k$ is not necessarily a descent direction. In this case, and more generally, when d_k and $\nabla\rho(z_k)$ are almost orthogonal, i.e.,

$$\cos\angle(d_k, \nabla\rho(z_k)) \leq \gamma \quad \text{with } \gamma \text{ small},$$

we reset the search direction d_k as the gradient

$$d_k = -\frac{\nabla\rho(z_k)}{\|\nabla\rho(z_k)\|_2}. \tag{26}$$

This safeguarding strategy ensures that the search direction d_k is descending, and it is gradient related (its orthogonal projection onto $-\nabla\rho(z_k)$ is uniformly bounded by a constant from below). Therefore, we can immediately conclude the global convergence of both Algs. 1 and 2 in the smooth situation. In particular, suppose $\mu^*(z)$ and $\xi^*(z)$ (hence also $\rho(z)$) are continuously differentiable in the level set $\{z\colon \rho(z) \leq \rho(z_0)\}$, then the iterates $\{z_k\}_{k=0}^{\infty}$ of both algorithms will be globally convergent to a stationary point z_*, namely, $\nabla\rho(z_*) = 0$. This result is a simple application of the standard global convergence analysis of line search methods using gradient related search directions; see, e.g. [1, Theorem 4.3].

Remark 3. Observe that, as long as $z_{k+1}^T H_k z_k \neq 0$, we can perform a line search with $d_k = z_{k+1} - z_k$. In that case, the first iteration in line 2 of Alg. 3 reduces to $\rho(z_k) - \rho(z_{k+1}) < -c d_k^T \nabla\rho(z_k)$. It is therefore only when the SCF iteration does not sufficiently decrease $\rho(z_{k+1})$ that we will apply the line search. In practice, we see that z_{k+1} from the SCF iteration usually leads to a reduced ratio $\rho(z_{k+1})$, so the line search is only applied exceptionally and it typically uses only 1 or 2 backtracking steps. This is in stark contrast to applying the steepest descent method directly to $\rho(z)$, where line search is used to determine the step size in each iteration.

5 Applications

In this section, we discuss three applications that give rise to the robust RQ optimization problem (3). We will in particular show that the closed-form formulations of the minimizers, as needed in (6), are indeed available and that the regularity assumption of Definition 1 is typically satisfied. This will allow us to apply our NEPv characterizations and SCF iterations from the previous sections. Numerical experiments for these applications are postponed to the next section.

5.1 Robust generalized eigenvalue classifier

Data classification via generalized eigenvalue classifiers can be described as follows. Let two classes of labeled data sets be represented by the rows of matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times n}$:

$$A = [a_1, a_2, \dots, a_m]^T \quad \text{and} \quad B = [b_1, b_2, \dots, b_p]^T,$$

where each row a_i and b_i is a point in the feature space $\mathbb{R}^n$ and $n \le \min\{m, p\}$. The generalized eigenvalue classifier (GEC), also known as multi-surface proximal support vector machine classification introduced in [20], determines two hyperplanes, one for each class, such that each plane will be as close as possible to one of the data sets and as far as possible from the other data set. Denoting the hyperplane related to class A as $w_A^T x - \gamma_A = 0$ with $w_A \in \mathbb{R}^n$ and $\gamma_A \in \mathbb{R}$, then (w_A, γ_A) needs to satisfy

$$\min_{\left[\begin{smallmatrix} w_A \\ \gamma_A \end{smallmatrix}\right] \neq 0} \frac{\|Aw_A - \gamma_A e\|_2^2}{\|Bw_A - \gamma_A e\|_2^2} = \min_{z=\left[\begin{smallmatrix} w_A \\ \gamma_A \end{smallmatrix}\right] \neq 0} \frac{z^T G z}{z^T H z}, \tag{27}$$

where e is a column vector of ones, $G = [A, -e]^T[A, -e]$ and $H = [B, -e]^T[B, -e]$. The other hyperplane (w_B, γ_B) for class B is determined similarly. Using these two hyperplanes, we predict the class label of an unknown sample $x_u \in \mathbb{R}^n$ by assigning it to the class with minimal distance from the corresponding hyperplane:

$$\text{class}(x_u) = \arg\min_{i\in\{A,B\}} \frac{|w_i^T x_u - \gamma_i|}{\|w_i\|}.$$

Observe that (27) is a generalized RQ minimization problem. We may assume that the matrices G and H are positive definite, since A and B are tall skinny matrices and the number of features n is usually much smaller than the number of points m and p. Hence, (27) is the minimal eigenvalue of the matrix pair (G, H).

To take into account uncertainty, we consider the data sets as $A + \Delta A$ and $B + \Delta B$ with $\Delta A \in \mathcal{U}_A$ and $\Delta B \in \mathcal{U}_B$, where $\mathcal{U}_A$ and $\mathcal{U}_B$ are the sets of admissible uncertainties. In [31], the authors consider each data point to be independently perturbed in the form of ellipsoids:

$$\mathcal{U}_A = \left\{ \Delta A = [\delta_1^{(A)}, \delta_2^{(A)}, \dots, \delta_m^{(A)}]^T \in \mathbb{R}^{m\times n} : \delta_i^{(A)T} \Sigma_i^{(A)} \delta_i^{(A)} \le 1, i = 1, \dots, m \right\}$$

and

$$\mathcal{U}_B = \left\{ \Delta B = [\delta_1^{(B)}, \delta_2^{(B)}, \dots, \delta_p^{(B)}]^T \in \mathbb{R}^{p\times n} : \delta_i^{(B)T} \Sigma_i^{(B)} \delta_i^{(B)} \le 1, i = 1, \dots, p \right\},$$

where $\Sigma_i^{(A)}, \Sigma_i^{(B)}$ are positive definite matrices defining the ellipsoids. The optimal hyperplane for the dataset A that takes into account the worst-case perturbations in the data points in (27) can be found by solving

$$\min_{z\neq 0} \max_{\substack{\Delta A \in \mathcal{U}_A \\ \Delta B \in \mathcal{U}_B}} \frac{z^T \widehat{G}(\Delta A) z}{z^T \widehat{H}(\Delta B) z}, \tag{28}$$

where $\widehat{G}(\Delta A) = [A+\Delta A,\ -e]^T[A+\Delta A,\ -e]$ and $\widehat{H}(\Delta B) = [B+\Delta B,\ -e]^T[B+\Delta B,\ -e]$. This is the robust RQ minimization problem (3). Since the inner maximization can be solved explicitly — see [31] and Appendix A.1 for a correction— it leads to the nonlinear RQ optimization problem (6) for the optimal robust hyperplane (w,γ), i.e.,

$$\min_{z=\left[\begin{smallmatrix}w\\ \gamma\end{smallmatrix}\right]\neq 0} \frac{z^TG(z)z}{z^TH(z)z}, \tag{29}$$

where the coefficient matrices are given by

$$G(z) = [A+\Delta A(z),\ -e]^T[A+\Delta A(z),\ -e], \tag{30a}$$

$$H(z) = [B+\Delta B(z),\ -e]^T[B+\Delta B(z),\ -e], \tag{30b}$$

with

$$\Delta A(z) = \begin{bmatrix} \frac{\operatorname{sgn}(w^Ta_1-\gamma)}{\sqrt{w^T\Sigma_1^{(A)-1}w}}\cdot w^T\Sigma_1^{(A)-1} \\ \frac{\operatorname{sgn}(w^Ta_2-\gamma)}{\sqrt{w^T\Sigma_2^{(A)-1}w}}\cdot w^T\Sigma_2^{(A)-1} \\ \vdots \\ \frac{\operatorname{sgn}(w^Ta_m-\gamma)}{\sqrt{w^T\Sigma_m^{(A)-1}w}}\cdot w^T\Sigma_m^{(A)-1} \end{bmatrix},\ \Delta B(z) = \begin{bmatrix} \varphi_1(z)\cdot\frac{\operatorname{sgn}(\gamma-w^Tb_1)}{\sqrt{w^T\Sigma_1^{(B)-1}w}}\cdot w^T\Sigma_1^{(B)-1} \\ \varphi_2(z)\cdot\frac{\operatorname{sgn}(\gamma-w^Tb_2)}{\sqrt{w^T\Sigma_2^{(B)-1}w}}\cdot w^T\Sigma_2^{(B)-1} \\ \vdots \\ \varphi_p(z)\cdot\frac{\operatorname{sgn}(\gamma-w^Tb_p)}{\sqrt{w^T\Sigma_p^{(B)-1}w}}\cdot w^T\Sigma_p^{(B)-1} \end{bmatrix}, \tag{31}$$

and

$$\varphi_j(z) = \min\left\{\frac{|\gamma-w^Tb_j|}{\sqrt{w^T\Sigma_j^{(B)-1}w}},\ 1\right\} \quad \text{for } j=1,2,\ldots,p.$$

The optimizer z_* of (29) defines the robust generalized eigenvalue classifier (RGEC) for the dataset A.

We note that the optimal parameter $\Delta A(z)$ is a function that is smooth in $z = \left[\begin{smallmatrix}w\\ \gamma\end{smallmatrix}\right] \in \mathbb{R}^{n+1}$ except for $z \in \bigcup_{i=1}^{m}\left\{\left[\begin{smallmatrix}w\\ \gamma\end{smallmatrix}\right] : w^Ta_i-\gamma=0\right\}$, i.e., when the hyperplane $\{x\colon w^Tx-\gamma=0\}$ defined by z touches one of the sampling points a_i. Likewise, $\Delta B(z)$ is smooth in z except for $z \in \bigcup_{j=1}^{p}\{z\colon \varphi_j(z)=1\}$, i.e., when the hyperplane defined by z is tangent to one of the ellipsoid of b_j. Since $\Delta A(z)$ and $\Delta B(z)$ represent the functions $\mu^*(z)$ and $\xi^*(z)$ in Definition 1, respectively, we can assume that for generic data sets, the optimal point z of (29) is regular.

5.2 Common spatial pattern analysis

Common spatial pattern (CSP) analysis is a technique commonly applied for feature extraction of electroencephalogram (EEG) data in brain-computer interface (BCI) systems, see for example [7]. The mathematical problem of interest is the RQ optimization problem

$$\min_{z\neq 0} \frac{z^T\overline{\Sigma}_-z}{z^T(\overline{\Sigma}_+ + \overline{\Sigma}_-)z}, \tag{32}$$

where $\overline{\Sigma}_+, \overline{\Sigma}_- \in \mathbb{R}^{n\times n}$ are symmetric positive definite matrices that are averaged covariance matrices of labeled signals $x(t) \in \mathbb{R}^n$ in conditions '$+$' and '$-$', respectively. By minimizing (32), we obtain a spatial filter z_+ for discrimination, with small variance for condition '$-$' and large variance for condition '$+$'. In analogy, we can also solve for the minimum to obtain the other spatial filter z_-. As a common practice in CSP analysis [7], the eigenvectors z_c are normalized such that $z_c^T(\overline{\Sigma}_+ + \overline{\Sigma}_-)z_c = 1$ for $c \in \{+,-\}$. These spatial filters are then used for extracting features.

Because of artifacts in collected data, the covariance matrices $\overline{\Sigma}_c$ can be very noisy, and it is important to robustify CSP against these uncertainties. In [13], the authors considered the robust CSP problem[3]

$$\min_{z\neq 0} \max_{\substack{\Sigma_+\in\mathcal{S}_+\\ \Sigma_-\in\mathcal{S}_-}} \frac{z^T\Sigma_- z}{z^T(\Sigma_+ + \Sigma_-)z}, \tag{33}$$

and the tolerance sets are given, for $c \in \{+,-\}$, by

$$\mathcal{S}_c = \left\{ \Sigma_c = \overline{\Sigma}_c + \Delta_c \;\middle|\; \Delta_c = \sum_{i=1}^{k} \alpha_c^{(i)} V_c^{(i)}, \quad \sum_{i=1}^{k} \frac{\left(\alpha_c^{(i)}\right)^2}{w_c^{(i)}} \leq \delta_c^2, \quad \alpha_c^{(i)} \in \mathbb{R} \right\}, \tag{34}$$

where δ_c are prescribed perturbation levels, $\overline{\Sigma}_c$ are nominal covariance matrices, $V_c^{(i)} \in \mathbb{R}^{n\times n}$ are symmetric interpolation matrices, and $w_c^{(i)}$ are weights for the coefficient variables $\alpha_c^{(i)}$. These parameters $\overline{\Sigma}_c$, $V_c^{(i)}$ and $w_c^{(i)}$ are typically obtained by principal component analysis of the signals; for more details, see numerical examples in §6

It is shown in [13] that the inner maximization problem in (33) can be solved explicitly, so (33) leads to the nonlinear RQ optimization problem

$$\min_{z\neq 0} \frac{z^T\Sigma_-(z)z}{z^T\left[\Sigma_+(z) + \Sigma_-(z)\right]z}, \tag{35}$$

where for $c \in \{+,-\}$,

$$\Sigma_c(z) = \overline{\Sigma}_c + \sum_{i=1}^{k} \alpha_c^{(i)}(z) V_c^{(i)} \quad \text{with} \quad \alpha_c^{(i)}(z) = \frac{-c\delta_c w_c^{(i)} z^T V_c^{(i)} z}{\sqrt{\sum_{i=1}^{k} w_c^{(i)} (z^T V_c^{(i)} z)^2}}. \tag{36}$$

Here, we have assumed $\Sigma_c(z) \succ 0$ which is guaranteed to hold if $\overline{\Sigma}_c$ is positive definite and if the perturbation levels δ_c are (sufficiently) small. In the general case when positive definiteness fails to hold, evaluating $\Sigma_c(z)$ can be done via some semi-definite programming (SDP) techniques. Observe that the optimal parameters $\alpha_c^{(i)}(z)$ are analytic functions of z. Hence, the optimal point z of (32) is regular in the sense of Definition 1 as long as the corresponding $\Sigma_c(z)$ is positive definite.

5.3 Robust Fisher linear discriminant analysis

Fisher's linear discriminant analysis (LDA) is widely used for pattern recognition and classification; see, e.g., [8]. Let $X, Y \in \mathbb{R}^n$ be two random variables, with mean μ_x and μ_y and covariance matrices Σ_x and Σ_y, respectively. LDA finds a discriminant vector $z \in \mathbb{R}^n$ such that the linear combination of variables z^TX and z^TY are best separated:

$$\max_{z\neq 0} \frac{(z^T\mu_x - z^T\mu_y)^2}{z^T\Sigma_x z + z^T\Sigma_y z}. \tag{37}$$

It is easy to see that up to a scale factor α, the optimal discriminant z_* is given by

$$z_* = \alpha \cdot (\Sigma_x + \Sigma_y)^{-1}(\mu_x - \mu_y).$$

[3] In [13], the problem is stated as $\max\limits_{z} \min\limits_{\Sigma_\pm} (z^T\Sigma_+ z)/(z^T(\Sigma_+ + \Sigma_-)z)$. To be consistent with our notation, we stated it into the equivalent min-max form (33) using $(z^T\Sigma_+ z)(z^T(\Sigma_+ + \overline{\Sigma}_-)z) = 1 - (z^T\Sigma_- z)(z^T(\Sigma_+ + \Sigma_-)z)$.

Consequently, the optimal z_* defines the LDA discriminant that can be used to generate the linear classifier $\varphi(u) = z_*^T u - \beta$ with $\beta = z_*^T(\mu_x + \mu_y)/2$. We can classify a new observation u by assigning it to class x if $\varphi(u) > 0$, and to class y otherwise.

In practice, the parameters μ_c and Σ_c for $c = \{x, y\}$ are unknown and replaced by their sample mean $\overline{\mu}_c$ and covariance $\overline{\Sigma}_c$, or similar estimates based on a finite number of samples. In any case, these quantities will be subject to uncertainty error which will influence the result of LDA. To account for these uncertainties, Kim, Magnani and Boyd [15] propose the robust LDA

$$\max_{z\neq 0} \min_{\substack{(\Sigma_x,\Sigma_y)\in\Omega \\ (\mu_x,\mu_y)\in\Gamma}} \frac{(z^T\mu_x - z^T\mu_y)^2}{z^T\Sigma_x z + z^T\Sigma_y z} \tag{38}$$

with the ellipsoidal uncertainty models

$$\begin{aligned} \Omega &= \{(\Sigma_x,\, \Sigma_y) :\ \|\Sigma_x - \overline{\Sigma}_x\|_F \leq \delta_x \text{ and } \|\Sigma_y - \overline{\Sigma}_y\|_F \leq \delta_y\}, \\ \Gamma &= \{(\mu_x,\, \mu_y) :\ (\mu_x - \overline{\mu}_x)^T S_x^{-1}(\mu_x - \overline{\mu}_x) \leq 1 \text{ and } (\mu_y - \overline{\mu}_y)^T S_y^{-1}(\mu_y - \overline{\mu}_y) \leq 1\}. \end{aligned} \tag{39}$$

Here, $\overline{\Sigma}_c \succ 0$, $\delta_c > 0$, and $S_c \succ 0$, $\overline{\mu}_c \in \mathbb{R}^n$ are known parameters for $c \in \{x, y\}$, but they can be estimated from the data.

To see that the robust LDA is of the robust RQ form (6), we write (37), by taking reciprocals, as the minimization of the RQ

$$\min_{z\neq 0} \frac{z^T(\Sigma_x + \Sigma_y)z}{z^T(\mu_x - \mu_y)(\mu_x - \mu_y)^T z}. \tag{40}$$

Therefore, the robust LDA in (38) is a solution of the robust RQ minimization

$$\min_{z\neq 0} \rho(z) \equiv \frac{\max\limits_{(\Sigma_x,\Sigma_y)\in\Omega} \quad z^T(\Sigma_x + \Sigma_y)z}{\min\limits_{(\mu_x,\mu_y)\in\Gamma} \quad z^T(\mu_x - \mu_y)(\mu_x - \mu_y)^T z}. \tag{41}$$

In Appendix A.2, we show that $\rho(z) = \infty$ if $|z^T(\overline{\mu}_x - \overline{\mu}_y)| \leq \sqrt{z^T S_x z} + \sqrt{z^T S_y z}$. Otherwise,

$$\rho(z) = \frac{z^T G z}{z^T H(z) z}, \tag{42}$$

where

$$G = \overline{\Sigma}_x + \overline{\Sigma}_y + (\delta_x + \delta_y) I_n \qquad \text{and} \qquad H(z) = f(z) f(z)^T,$$

and

$$f(z) = (\overline{\mu}_x - \overline{\mu}_y) - \operatorname{sgn}(z^T(\overline{\mu}_x - \overline{\mu}_y)) \left(\frac{S_x z}{\sqrt{z^T S_x z}} + \frac{S_y z}{\sqrt{z^T S_y z}} \right).$$

Note that $H(z)$ is analytic in z when $z^T(\overline{\mu}_x - \overline{\mu}_y) \neq 0$, so z is regular in the sense of Definition 1 if $\rho(z) < \infty$.

In [15, 14], it is shown that the max-min problem (38) satisfies a saddle-point property that allows us to formulate it as the equivalent problem

$$\begin{aligned} \min \quad & (\mu_x - \mu_y)^T G^{-1} (\mu_x - \mu_y) \\ \text{s.t.} \quad & (\mu_x, \mu_y) \in \Gamma. \end{aligned} \tag{43}$$

With the constraints (39), the optimization problem (43) is a well-studied convex quadratically constrained program (QCP). The global minimizer (μ_x^*, μ_y^*) can be solved, e.g., by CVX, a package for specifying and solving convex programs [11]. The global minimizer of (42) is given by $z_* = G^{-1}(\mu_x^* - \mu_y^*)$.

Despite that (43) can be formulated as a QCP, it may be more beneficial to use the NEPv characterization described in the previous section. Indeed, as shown by numerical experiments in §6, Alg. 2 usually converges in less than 10 iterations and is more accurate since it avoids the construction of explicit inverses, i.e., there is no need for G^{-1} and S_c^{-1}. Moreover, the SCF iterations always converged to a (numerically) global minimum. This rather surprising property can be explained from the following result.

Theorem 4. *Suppose $\rho(z) \not\equiv \infty$, then any eigenpair of the NEPvs (12), (14) or (20) is a global solution of the robust LDA problem (38).*

Proof. It is sufficient to show the theorem for the NEPv (12). We first recall that the robust problem (38) is equivalent to the convex program (43), where the sufficient and necessary condition for global optimality is given by the KKT condition

$$\begin{array}{lll} \nu > 0, & G^{-1}(\mu_x - \mu_y) + \nu S_x^{-1}(\mu_x - \overline{\mu}_x) = 0, & (\mu_x - \overline{\mu}_x)^T S_x^{-1}(\mu_x - \overline{\mu}_x) - 1 = 0, \\ \gamma > 0, & G^{-1}(\mu_x - \mu_y) - \gamma S_y^{-1}(\mu_y - \overline{\mu}_y) = 0, & (\mu_y - \overline{\mu}_y)^T S_y^{-1}(\mu_y - \overline{\mu}_y) - 1 = 0, \end{array} \tag{44}$$

where we exploited $\nu \neq 0$ and $\gamma \neq 0$ since otherwise $\mu_x = \mu_y$ which implies that the denominator in $\rho(z)$ in (41) vanishes and $\rho(z) \equiv \infty$ for all $z \neq 0$.

Next, let (λ, z) be an eigenpair of the NEPv (12), it holds $\lambda = \rho(z) > 0$. Due to homogeneity of $f(z)$, we can take z such that $\lambda f(z)^T z = 1$ and obtain

$$Gz = \lambda(f(z)f(z)^T)z = f(z). \tag{45}$$

Since $f(z)^T z > 0$, it follows from definition (42) and the condition $\rho(z) < \infty$ that

$$\sigma := \operatorname{sgn}(z^T(\overline{\mu}_x - \overline{\mu}_y)) = 1.$$

Now observe that

$$\mu_x = -\frac{\sigma S_x z}{\sqrt{z^T S_x z}} + \overline{\mu}_x, \quad \mu_y = \frac{\sigma S_y z}{\sqrt{z^T S_y z}} + \overline{\mu}_y, \quad \nu = \frac{\sqrt{z^T S_x z}}{\sigma}, \quad \gamma = \frac{\sqrt{z^T S_y z}}{\sigma}$$

satisfy (44). Therefore, (μ_x, μ_y) a global minimizer of (43) and so the eigenvector $z = G^{-1}f(z) = G^{-1}(\mu_x - \mu_y)$ is a global minimizer of $\rho(z)$. □

6 Numerical examples

We report on numerical experiments for the three applications discussed in §5. We will show the importance of the eigenvalue ordering issue for the derived NEPv, and the potential divergence problem of the simple iterative scheme (7). All computations were done in MATLAB 2017a. The starting values z_0 for all algorithms are set to the nonrobust minimizer, i.e., solution of (27) for robust GEC, of (32) for robust CSP, and of (40) for robust LDA, respectively. The tolerance for both Algs. 1 and 2 is set to `tol` $= 10^{-8}$. Safeguarded line search is also applied. In the spirit of reproducible research, all MATLAB scripts and data that were used to generate our numerical results can be found at http://www.unige.ch/~dlu/.

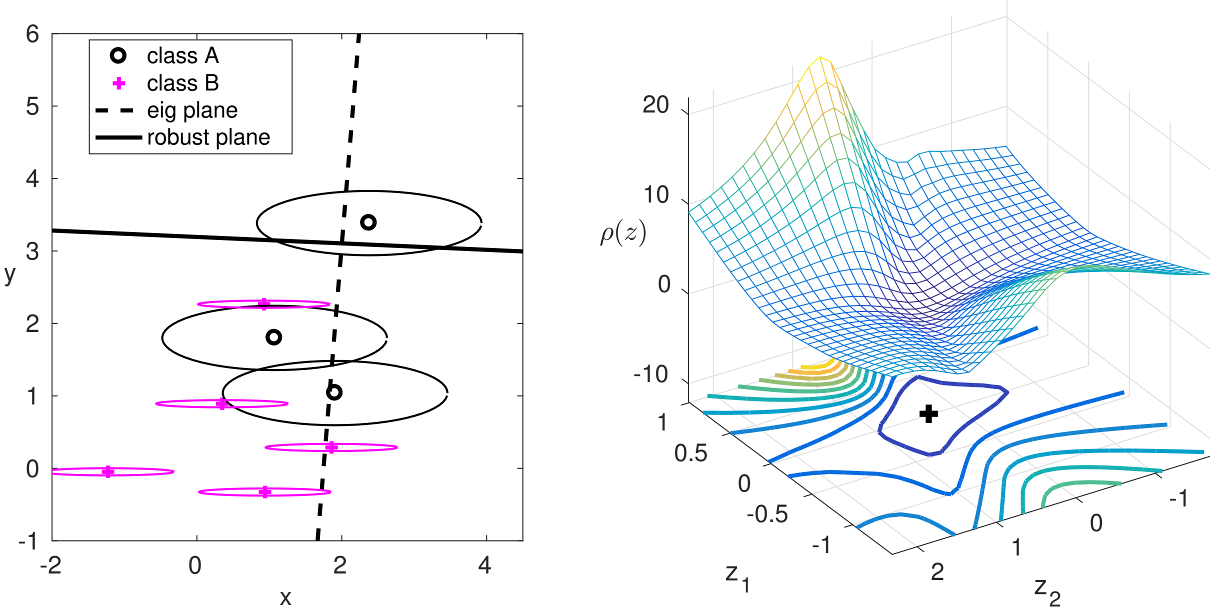

Figure 1: Left: data sets A and B, uncertainty and eigenclassifiers. Right: magnitudes of $\rho(z)$ close to local minimizer $\widehat{z}_*$ (marked as **+**).

Example 1 (Robust GEC)**.** In this example, we use a synthetic example to illustrate the convergence behavior of SCF iterations, and the associated the eigenvalue-ordering issue. Let data points $a_i, b_i \in \mathbb{R}^2$ be chosen as shown in the left plot of Fig. 1, together with their uncertainty ellipsoids. GEC for the dataset A (without robustness consideration) computed by the RQ optimization (27) is shown by the dashed line, and RGEC is shown by the solid line. The minimizer computed by Alg. 1 or 2 for RGEC is

$$\widehat{z}_* \approx \begin{bmatrix} 0.013890 & 0.313252 & 1.0 \end{bmatrix} \quad \text{with} \quad \rho(\widehat{z}_*) \approx 0.2866130.$$

We can see that RGEC faithfully reflects the trend of uncertainties in dataset A. Note that since RGEC represented by $\widehat{z}_*$ does not pass through any of the data points, the solution $\widehat{z}_*$ is a regular point. To examine the local minimum, the right plot in Fig. 1 shows the magnitude of $\rho(z)$ for $z = [z_1, z_2, 1]$ close to $\widehat{z}_*$. Note that since $\rho(z) = \rho(\alpha z)$ for $\alpha \neq 0$, we can fix the coordinate $z_3 = 1$.

The convergence behavior of the robust Rayleigh quotients $\rho(z_k)$ by three different SCF iterations is depicted in Fig. 2. Alg. 2 for the second-order NEPv (14) shows superlinear convergence as proven in Theorem 3. Alg. 1 for the first-order NEPv (12) rapidly reaches a moderate accuracy of about 10^{-3} but it only converges linearly. We also see that the simple iterative scheme (7), which is proposed in [31, Alg. 1], fails to converge.

Let us check the eigenvalue order of the computed eigenpair $(\rho(\widehat{z}_*), \widehat{z}_*)$. The first three eigenvalues at $\widehat{z}_*$ of the first-order and second-order NEPv (12) and (14) are given by

$$\begin{array}{llll} \text{First-order NEPv (12):} & \lambda_1 = 0.1328986 & \underline{\lambda_2 = 0.2866130} & \lambda_3 = 2.8923953 \\ \text{Second-order NEPv (14):} & \lambda_1 = -0.2946578 & \underline{\lambda_2 = 0.2866130} & \lambda_3 = 2.8433553 \end{array}$$

We can see that the minimal ratio $\rho(\widehat{z}_*)$ is the least positive eigenvalue of the second-order NEPv (14) but it is not the smallest eigenvalue of the first-order NEPv (12). As explained in §4.1, we cannot therefore expect that the simple iterative scheme (7) converges to $\widehat{z}_*$. In addition,

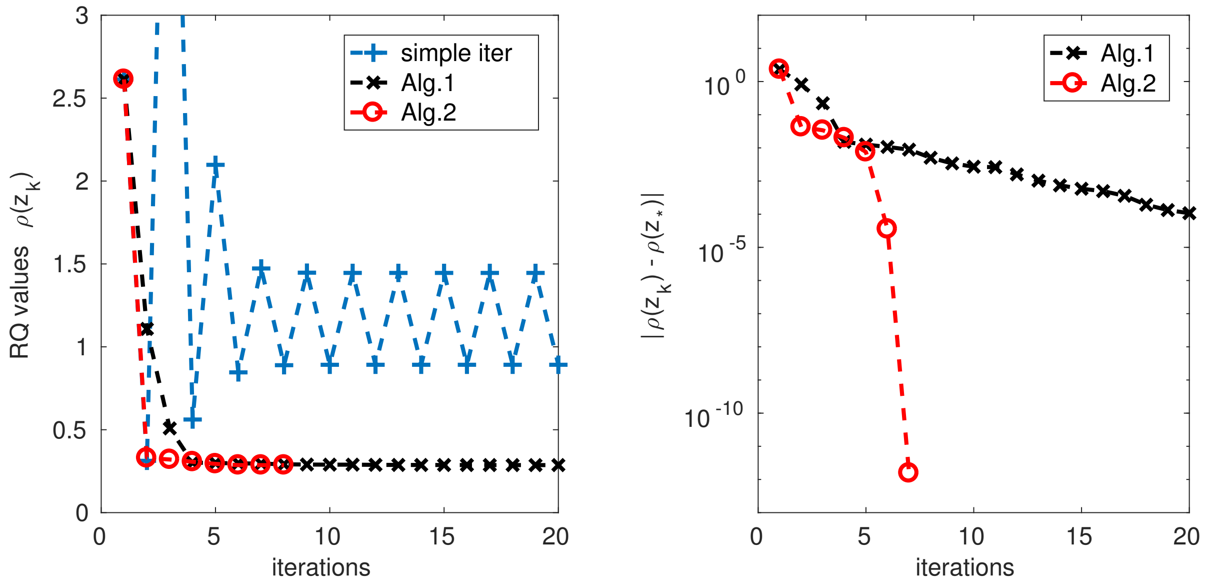

Figure 2: Left: convergence behaviors of three SCF iterations. Right: errors of $\rho(z_k)$.

from the convergence behavior in the left plot of Fig. 2, we see that the objective value $\rho(z_k)$ oscillates between two points that are neither an optimal solution. This shows the usefulness of the nonlinear spectral transformation in Alg. 1.

Example 2 (Robust GEC)**.** In this example, we apply RGEC to the Pima Indians Diabetes (PID) dataset [26] in the UCI machine learning repository [18]. In this dataset, there are 768 data points classified into 2 classes (diabetes or not). Each data point x collects 8 attributes (features) such as blood pressure, age and body mass index (BMI) of a patient. For numerical experiments, we set the uncertainty ellipsoid for each patient data to be of the form

$$\Sigma^{-1} = \operatorname{diag}(\alpha_1^2\bar{x}_1^2,\, \alpha_2^2\bar{x}_2^2,\, \dots,\, \alpha_n^2\bar{x}_n^2). \tag{46}$$

where $\bar{x}_i$ is the mean of the ith feature x_i over all patients, and α_i is a measure for the anticipated relative error of x_i. We set $\alpha_i = 0.5$ (hence, 50% relative error) for all features, except for the 1st (number of times of pregnant) and the 8th (age), where we set $\alpha_i = 0.001$ since we do not expect large errors in those features.

Similar to the setup in [31], we apply holdout cross validation with 10 repetitions. In every repetition, 70% of the randomly chosen data points are used as *training set* and the remaining 30% as *testing set*. The training set is used to compute the two classification planes, given the uncertainty ellipsoid Σ if required. Testing is performed by classifying random data points $x + \delta x$ with x a sample from the testing set and $\delta x \sim \mathcal{N}(0, \beta\Sigma)$, a normal distribution with mean 0 and variance $\beta\Sigma$. Each sample in the testing set is used exactly once. The factor $\beta > 0$ expresses the conservativeness of the uncertainty ellipsoid. Since δx is normally distributed, a sample $x + \delta x$ is more likely to violate the ellipsoidal constraints with growing β. We will use the following values in the experiments: $\beta = 0.1, 1, 10$. For each instance of the training set, we perform 100 such classification tests and calculate the best, worst, and average classification accuracy (ratio of number of correctly classified samples to the total number of tests).

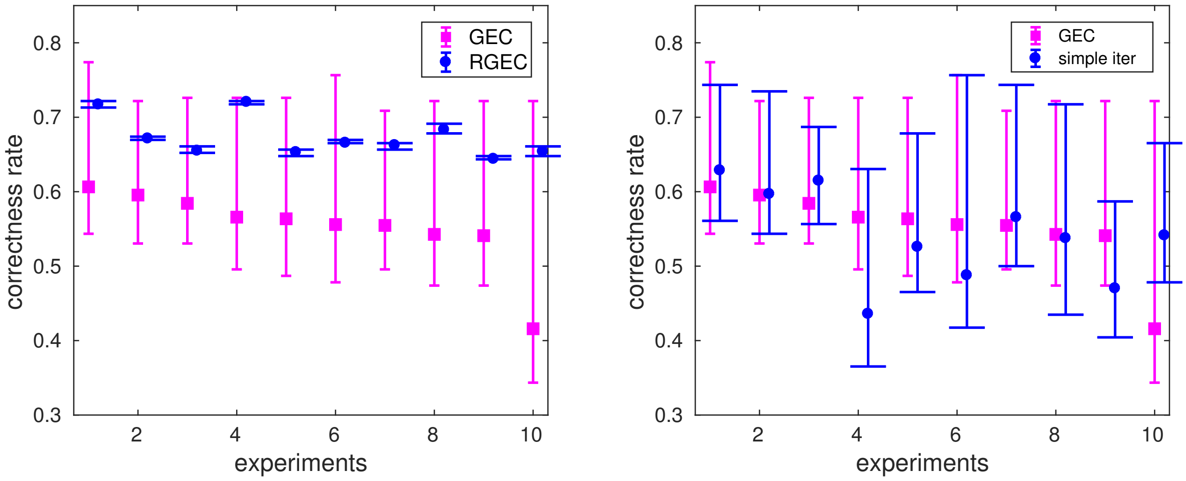

Figure 3: Correctness rates for the PID dataset with $\beta = 0.1$. The squares and dots represent the average rates, and the error bars depict the range between the best and worst rate. The experiments are sorted in decreasing order of average rates for GEC. Left: GEC in magenta and RGEC in blue. Right: GEC in magenta and the simple iterative scheme (7) in blue.

At the first experiment with $\beta = 0.1$, we compare GEC and RGEC. We observed convergence in all experiments. We summarize the correctness rates of the classification in left plot of Fig. 3. RGEC shows very small variance. In contrast, GEC demonstrates large variance, and lower average correctness rates. For comparison, we also reported the testing results (on the same data) for the simple iterative scheme (7) in the right plot of Fig. 3. Since the simple iteration does not always converge, we took the solution with the smallest nonlinear RQ within 30 iterations. The results for the other values of β are reported in Fig. 4. As β increases, RGEC significantly improves the results of GEC.

Example 3 (Robust CSP)**.** We consider a synthetic example of CSP analysis discussed in §5.2. As described in [13], the testing signals are generated by a linear mixing model with nonstationary sources:

$$x(t) = A \begin{bmatrix} s^d(t) \\ s^n(t) \end{bmatrix} + \varepsilon(t),$$

where $x(t)$ is a 10-dimensional signal, $s^d(t)$ is a 2-dimensional discriminative source, $s^n(t)$ is an 8-dimensional non-discriminative source, A is a random rotation, and $\varepsilon(t) \sim \mathcal{N}(0, 2)$. The discriminative source $s^d(t)$ is sampled from $\mathcal{N}\left(0, \mathrm{diag}(1.8,\, 0.6)\right)$ in condition '+', and $\mathcal{N}\left(0, \mathrm{diag}(0.2,\, 1.4)\right)$ in condition '−'. The non-discriminative sources $s^n(t)$ are sampled from $\mathcal{N}(0, 1)$ in both conditions. For each condition $c \in \{+, -\}$, we generate $N = 50$ random signals $x(t)$ that are sampled in $m = 200$ points to obtain the matrix $X_c^{(j)} = [x(t_1), x(t_2), \ldots, x(t_m)]$ for $j = 1, \ldots, N$.

To obtain the coefficients $\overline{\Sigma}_c$, $V_c^{(i)}$ and $w_c^{(i)}$ in the tolerance sets $\mathcal{S}_c$ (34), we apply the PCA scheme described in [13]. In particular, for each condition $c \in \{+, -\}$, we first compute the (local) covariance matrix $U_c^{(j)} = \frac{1}{m-1} \sum_{i=1}^{m} X_c^{(j)}(:, i) X_c^{(j)}(:, i)^T$ for $j = 1, \ldots, N$, and define $\overline{\Sigma}_c = \frac{1}{N} \sum_{j=1}^{N} U_c^{(j)}$ as the averaged covariance matrix. We then vectorize each $(U_c^{(j)} - \overline{\Sigma}_c) \in \mathbb{R}^{10\times 10}$ to

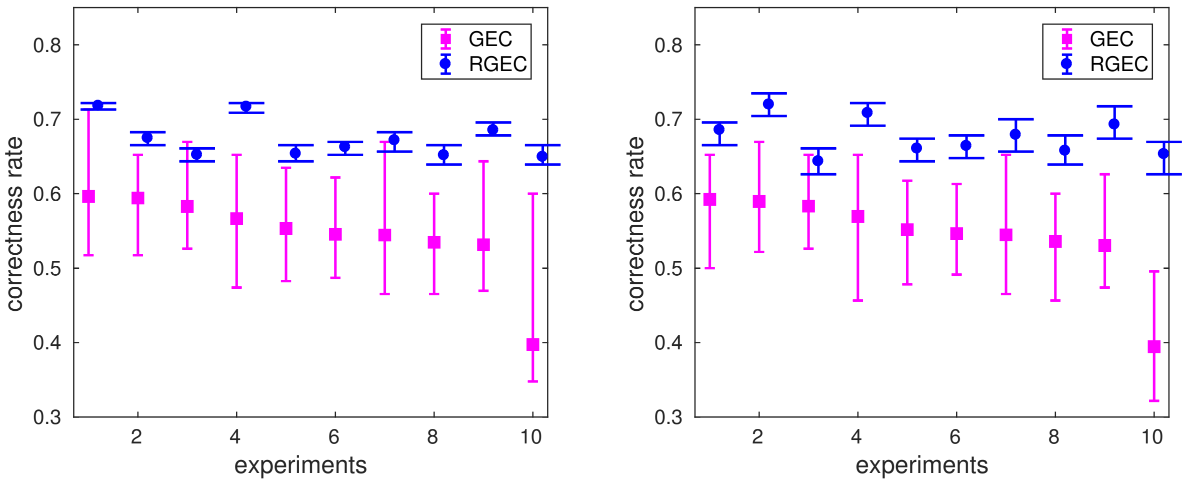

Figure 4: Correctness rates for the PID dataset with $\beta = 1$ (left) and $\beta = 10$ (right).

$u_c^{(j)} \in \mathbb{R}^{100}$, and compute the singular value decomposition

$$[u_c^{(1)}, u_c^{(2)}, \dots, u_c^{(N)}] = \sum_{i=1}^{N} \sigma_c^{(i)} v_c^{(i)} q_c^{(i)T},$$

where $\sigma_c^{(1)} \geq \cdots \geq \sigma_c^{(N)} \geq 0$ are ordered singular values, and $\{v_c^{(i)}\}_{i=1}^N \in \mathbb{R}^{100}$ and $\{q_c^{(i)}\}_{i=1}^N \in \mathbb{R}^N$ the corresponding left and right singular vectors. For numerical experiments, we take the leading $k = 10$ singular values to define $w_c^{(i)} = (\sigma_c^{(i)})^2$, matricize (inverse of vectorization) the singular vectors $v_c^{(i)} \in \mathbb{R}^{100}$ to $V_c^{(i)} \in \mathbb{R}^{10\times 10}$, and symmetrize $V_c^{(i)} := (V_c^{(i)} + V_c^{(i)T})/2$, for $i = 1, \dots, k$.

To show the convergence of SCF iterations, we compute the minimizer of the nonlinear RQ (35) with perturbation $\delta_+ = \delta_- = 6$. Both Algs. 1 and 2 converge to a (local) optimal value $\rho(\widehat{z}_*) = 1.042032$. Some ordered eigenvalues at $\widehat{z}_*$ are listed below

First-order NEPv:	$\cdots$	$\lambda_5 = 1.017731$	$\underline{\lambda_6 = 1.042032}$	$\lambda_7 = 1.239586$
Second-order NEPv:	$\cdots$	$\lambda_2 = -0.527799$	$\underline{\lambda_3 = 1.042032}$	$\lambda_4 = 1.286031$

The largest eigenvalue of the first-order NEPv (12) is $\lambda_{10} \approx 2.7$ from which we compute $\sigma(z)$ for the shift. The optimal $\rho(\widehat{z}_*)$ corresponds to the least positive eigenvalue of the second-order NEPv (14), and the 6th eigenvalue of the first-order NEPv (12). In the convergence plot of Fig. 5, we see that the simple iterative scheme (7) (used as [13, Alg. 1] to solve (35)) fails to converge. Alg. 2 is locally quadratically convergent. Alg. 1 converges quickly in the first few iterations. This shows the potential of combining Algs. 1 and 2 for fast global convergence.

Example 4 (Robust CSP)**.** In this example, we use the computed spatial filters z_+ and z_- for signal classification as in BCI systems. To predict the class label of a sampled signal $X = [x(t_1), x(t_2), \dots, x(t_m)]$, a common practice in CSP analysis (see, e.g., [7]) is to first extract the log variance feature of the signal using the spatial filters

$$f(X) = \log\left(\begin{bmatrix}\mathrm{var}(z_+^T X)\\ \mathrm{var}(z_-^T X)\end{bmatrix}\right),$$

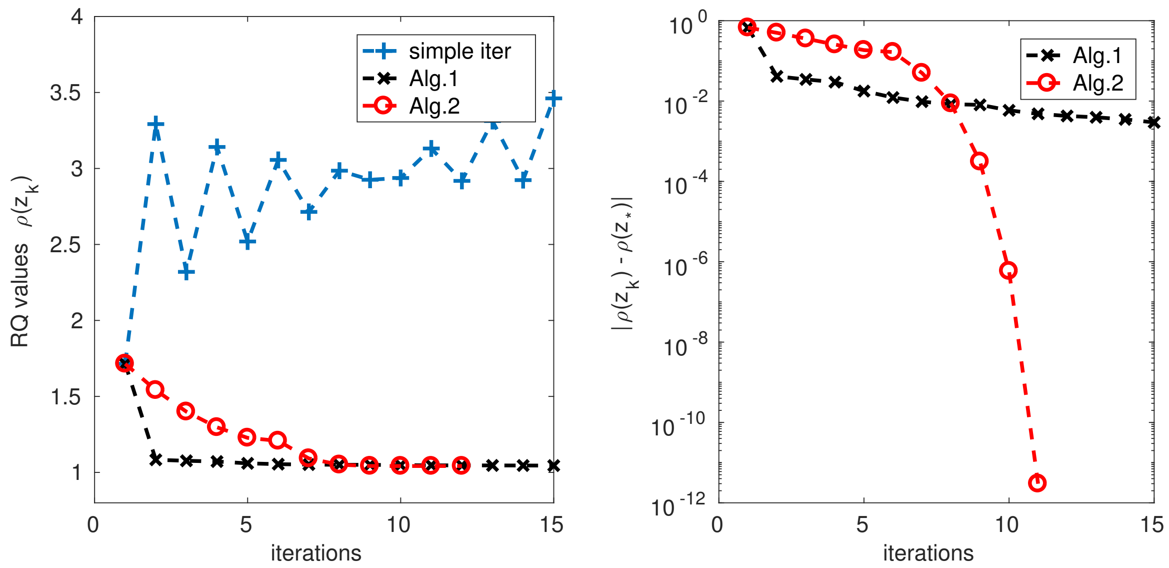

Figure 5: Convergence of $\rho(z_k)$ of CSP analysis, synthetic example

where the variance $\mathrm{var}(x) := \sum_{i=1}^{m}(x_i - \mu)^2/(m-1)$ and the mean $\mu = \sum_{i=1}^{m} x_i/m$ and the elementwise logarithm $\log(\cdot)$, and then define a linear classifier

$$\varphi(X) = w^T f(X) - \beta_0, \tag{47}$$

where β_0 and $w \in \mathbb{R}^2$ are weights. The sign of $\varphi(X)$ is used for the class label of signal X.

The weights w and β_0 are determined by training signals using Fisher's linear discriminant analysis (LDA) (see, e.g., [8]). Specifically, let $f_c^{(i)} = f(X_c^{(i)})$ be the log variance features of the training signals for $i = 1, \ldots, N$ and

$$S_c = \sum_{i=1}^{N}(f_c^{(i)} - m_c)(f_c^{(i)} - m_c)^T \quad \text{with} \quad m_c = \frac{1}{N}\sum_{i=1}^{N} f_c^{(i)}$$

be the corresponding scatter matrices, where $c \in \{+,-\}$, then the weights w and β_0 are determined by $w = \widetilde{w}/\|\widetilde{w}\|_2$ with $\widetilde{w} = (S_+ + S_-)^{-1}(m_+ - m_-)$, and $\beta_0 = \frac{1}{2}w^T(m_+ + m_-)$.

For numerical experiments, we train the classifier (47) using the synthetic signals from Example 3. The spatial filters z_+ and z_- are computed from either CSP, i.e., using averaged covariance matrices $\overline{\Sigma}_+$ and $\overline{\Sigma}_-$, or robust CSP, i.e., using Alg. 2 with $\delta_c = 0.5, 1, 2, 4, 6, 8$, for $c \in \{+,-\}$. To assess the classifiers under uncertainties, we generate and classify a test signal from the same linear model but with an increased noise term $\varepsilon(t) \sim \mathcal{N}(0, 30)$. We repeated the experiment 100 times and summarize the results in Fig. 6. We observe significant improvements of the classification correctness rates for robust CSP with properly chosen perturbation levels δ. The choice of δ is clearly critical for the performance (as also discussed in [13]) but a good value can be estimated in practice by cross validation. For comparison, we also reported the results for the simple iterative scheme (7), where the solution with the smallest $\rho(z_k)$ is retained in case of non-convergence. In Fig. 7, the same experiment is repeated but now with noise terms $\varepsilon(t) \sim \mathcal{N}(0, 10)$ and $\varepsilon(t) \sim \mathcal{N}(0, 20)$. For both the robust and the non-robust algorithms, the classification rates improve as expected with smaller noise. However, robust CSP still gives considerably better results showing the robustness of our approach to the magnitude of the noise.

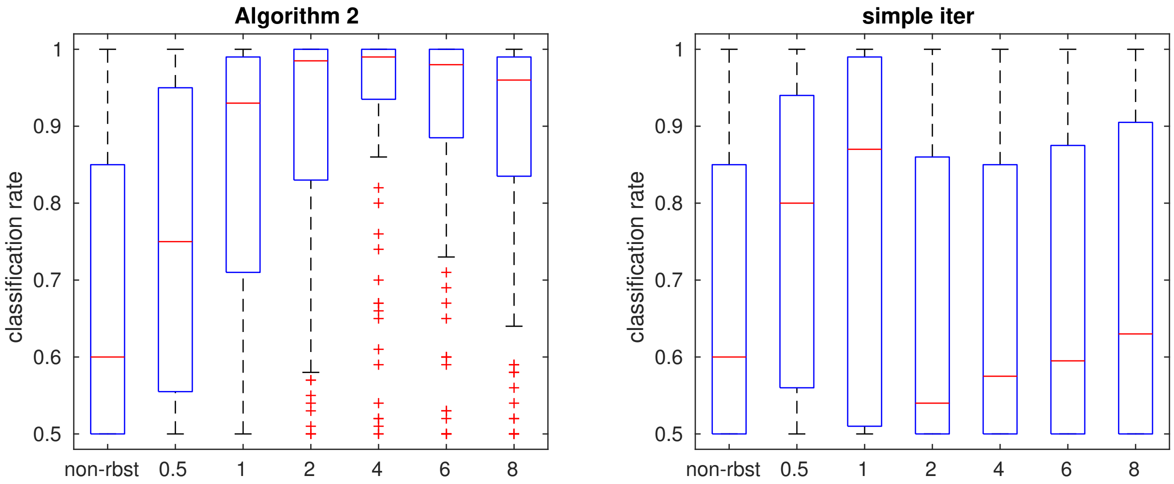

Figure 6: The boxplot of the classification rate for the linear mixing model problem with $\varepsilon(t) \sim \mathcal{N}(0, 30)$. The boxes from left to right represent standard CSP (non-rbst), and robust CSP with $\delta \in \{0.5, 1, 2, 4, 6, 8\}$. The robust CSP is computed by Alg. 2 (left panel) and the simple iterative scheme (right panel).

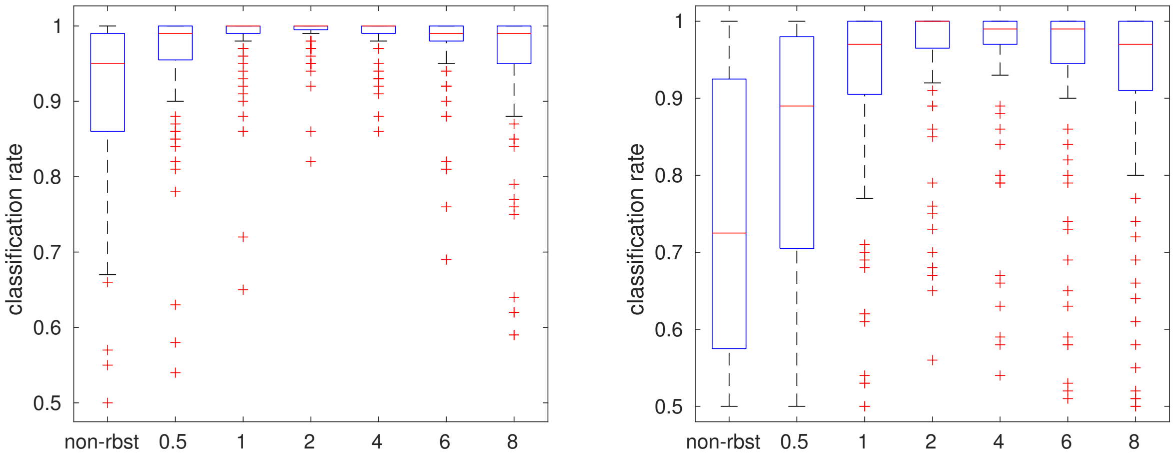

Figure 7: The classification rate of robust CSP (by Alg. 2) for the linear mixing model problem with $\varepsilon(t) \sim \mathcal{N}(0, 10)$ (left) and $\mathcal{N}(0, 20)$ (right).

Example 5 (Robust LDA)**.** In this example we demonstrate the effectiveness of NEPv approach for solving the robust LDA problems from §5.3. We use the sonar and ionosphere benchmark problems from the UCI machine learning repository [18]. The sonar problem has 208 points each with 60 features, and ionosphere has 351 points each with 34 features. Both benchmark problems are used in [15] for testing robust LDA, and we will follow the same setup here.

For the experiment, we randomly partition the data set into training and testing sets. The number of training points to the total is controlled by a ratio α. For a given partition, we generate the uncertainty parameters (39) by resampling technique. In particular, we resampled the training set with uniform distribution over all data points and then compute the sample mean and covariance matrices of each data class for the resampled training set. We repeat this 100 times. The averaged covariance matrices are used to define $\overline{\Sigma}_x$, and $\overline{\Sigma}_y$, whereas the maximum deviation (in Frobenius norm) to the average is used to define δ_x and δ_y, respectively. In the same fashion, the averaged mean values are used to define $\overline{\mu}_x$ and $\overline{\mu}_y$, whereas the covariance of all the mean values, P_x and P_y, are used to define $S_x = nP_x$ and $S_y = nP_y$.

Using these uncertainty parameters for (39), we compute the robust discriminant and evaluate the classification accuracy for the testing set. We repeat such classification experiment 100 times (each time with a new random partition), and obtain the average accuracy and the deviation. In Fig. 8 we reported the results for various partition parameters α. The robust discriminants are computed by Alg. 2. Fig. 8 reproduces the results demonstrated in [15]. It shows that RLDA significantly improves the classification accuracy over the conventional LDA (using averaged mean and covariance matrices).

In our experiments, Alg. 2 successfully find the minimizers for all robust RQs (100 × 6 cases for each data set) with the specified tolerance $\texttt{tol} = 10^{-8}$. It also showed fast convergence: The average number of iterations (and hence linear eigenvalue problems) was 8.79 for the ionosphere problem and 8.01 for the sonar problem. The overall computation time was 2.8 and 7.2 seconds, respectively. For comparison, when the robust LDA problem is solved as a QCP with CVX [11], the overall computation time was 136.9 and 163.4 seconds, respectively. We can also alternatively solve the first-order NEPv by Alg. 1. That will produce the same results, but with a larger number of iterations to reach high accuracy. Observe that since the matrix pair $(G, H(z))$ has only one positive eigenvalue, there is no need to reorder the eigenvalues in NEPv (12).

We remark that both Algs. 1 and 2 have to start with z_0 s.t. $\rho(z_0) \neq \infty$. In our experiment, this was not a problem since the non-robust solution always provided a valid z_0. However, if $\rho(z_0) = \infty$ happens, then one has to reset z_0 by checking the feasibility of $|z^T(\overline{\mu}_x - \overline{\mu}_y)| > \sqrt{z^TS_xz} + \sqrt{z^TS_yz}$ for $z \neq 0$, i.e., the two ellipsoids Γ in (39) do not intersect. This can be done with convex optimization.

7 Concluding remarks

We introduced the robust RQ minimization problem and reformulated it to nonlinear eigenvalue problems with eigenvector nonlinearity (NEPv). Two forms of NEPv were derived, namely one that only uses first-order information, while the other also uses second-order derivatives. Attention was paid to the eigenvalue ordering issue in solving the nonlinear eigenvalue problem via self-consistent field (SCF) iterations that may lead to non-convergence. To solve the eigenvalue ordering issue, we introduced a nonlinear spectral transformation technique for the first-order NEPv. The SCF iteration for the second-order NEPv has proven local quadratic convergence. The effectiveness of the proposed approaches are demonstrated by numerical experiments arising in three applications from data science.

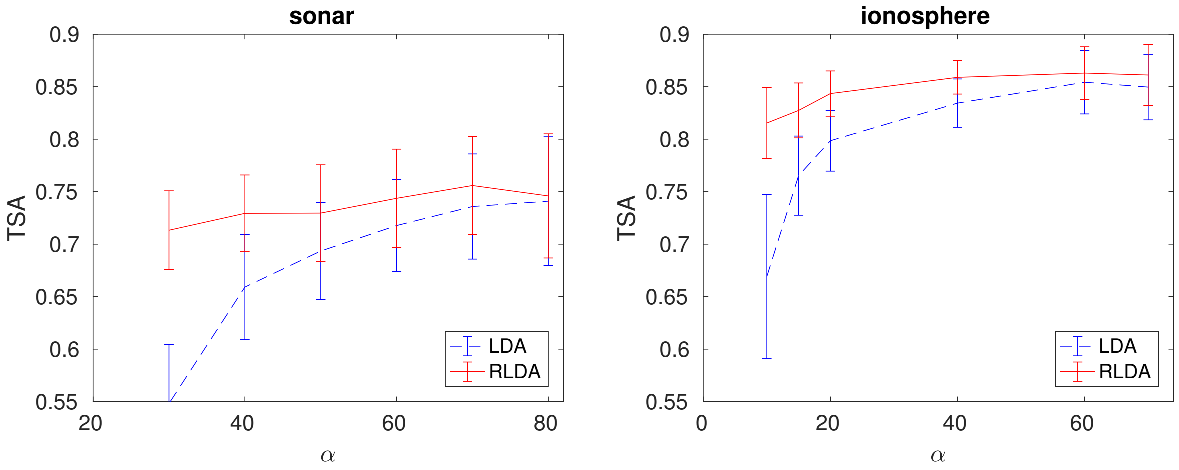

Figure 8: Test set accuracy (TSA) for sonar and ionosphere benchmark problems. The robust discriminant of RLDA is computed by Alg. 2.

The results presented in this work depend on the smoothness assumption of the optimal parameters $\mu^*(z)$ and $\xi^*(z)$. The smoothness condition allows us to employ the nonlinear eigenvalue characterization in Theorems 1 and 2, and consequently the SCF iterations in Algs. 1 and 2 can be applied. This assumption is satisfied for the applications discussed in this paper, however, it is a subject of future study on how to solve the robust RQ minimization when this assumption does not hold.

A Proofs related to equivalent formulations

A.1 Inner minimization for robust eigenclassifier

The following lemma provides a correction for a similar result in [31]. The difference is in the use of the function $\varphi(w)$.

Lemma 4. *Given vectors $w \neq 0$ and $x_c \in \mathbb{R}^n$, a symmetric positive definite matrix $\Sigma \in \mathbb{R}^{n\times n}$, and scalar γ, the following holds:*

(a) The maximization problem satisfies

$$\max_{x^T\Sigma x\leq 1} \left(w^T(x_c+x)-\gamma\right)^2 = \left(w^T(x_c+x_*)-\gamma\right)^2, \tag{48}$$

where

$$x_* = \frac{sgn(w^Tx_c-\gamma)}{\sqrt{w^T\Sigma^{-1}w}}\Sigma^{-1}w.$$

(b) The minimization problem satisfies

$$\min_{x^T\Sigma x\leq 1} \left(w^T(x_c+x)-\gamma\right)^2 = \left(w^T(x_c+x_*)-\gamma\right)^2, \tag{49}$$

where

$$x_* = \varphi(\gamma,w)\cdot\frac{sgn(\gamma-w^Tx_c)}{\sqrt{w^T\Sigma^{-1}w}}\Sigma^{-1}w \quad and \quad \varphi(\gamma,w)=\min\left\{\frac{|\gamma-w^Tx_c|}{\sqrt{w^T\Sigma^{-1}w}},\ 1\right\}.$$

The function $\varphi(\gamma, w)$ is smooth except for $|\gamma - w^T x_c| = \sqrt{w^T \Sigma^{-1} w}$, i.e., the hyperplane $\{x \colon w^T(x_c + x) - \gamma\}$ is tangent to the ellipsoid $x^T \Sigma^{-1} x = 1$.

Proof. (a) Let us define the Lagrangian of the maximization problem

$$L(x, \lambda) = \big(w^T(x_c + x) - \gamma\big)^2 - \lambda(x^T \Sigma x - 1),$$

where λ is the Lagrangian multiplier. The maximum (x_*, λ_*) must satisfy the KKT conditions

$$\text{stationary:} \quad L_x(x_*, \lambda_*) := 2\big(w^T(x_c + x_*) - \gamma\big)w - 2\lambda_* \Sigma x_* = 0 \tag{50}$$

$$\text{feasibility:} \quad \lambda_* \geq 0 \quad \text{and} \quad x_*^T \Sigma x_* \leq 1 \tag{51}$$

$$\text{slackness:} \quad \lambda_* \cdot (x_*^T \Sigma x_* - 1) = 0. \tag{52}$$

The multiplier $\lambda_* > 0$ must be strictly positive, since otherwise $\lambda_* = 0$ and the stationary condition (50) implies the maximum of (48) $(w^T(x_c + x_*) - \gamma)^2 = 0$ so the ellipsoid $x^T \Sigma x \leq 1$ degenerates to a plane. The positivity of λ_*, combined with condition (50), implies $x_* = \alpha \Sigma^{-1} w$ with α being a scalar. Plugging x_* into the slack condition (52), we obtain $\alpha = \pm(w^T \Sigma^{-1} w)^{-1/2}$, i.e.,

$$x_* = \pm(w^T \Sigma^{-1} w)^{-1/2} \cdot \Sigma^{-1} w.$$

We choose the sign of the leading coefficient that maximizes the optimizing function (48) at x_*, and obtain the expression in (48).

(b) First, suppose the intersection of the ellipsoid and the hyperplane

$$\mathcal{S} := \big\{x \colon x^T \Sigma x \leq 1\big\} \bigcap \big\{x \colon \ w^T(x_c + x) - \gamma = 0\big\} \ = \emptyset, \tag{53}$$

then the minimization problem

$$\min_{x^T \Sigma x \leq 1} \big(w^T(x_c + x) - \gamma\big)^2 = \big(w^T(x_c + x_*) - \gamma\big)^2, \tag{54}$$

where

$$x_* = \frac{\operatorname{sgn}(\gamma - w^T x_c)}{\sqrt{w^T \Sigma^{-1} w}} \Sigma^{-1} w.$$

The proof is analogous to Lemma 4 except for $\lambda_* \leq 0$ in the feasibility condition (51) due to the minimization. The nonvanishing condition (53) ensures the corresponding multiplier $\lambda_* < 0$ is strictly negative, since otherwise $\lambda_* = 0$ leads to $\big(w^T(x_c + x_*) - \gamma\big)w = 0$ with $x_*^T \Sigma x_* \leq 1$, contradicting $\mathcal{S} = \emptyset$.

If $\mathcal{S}$ is nonempty, i.e.,

$$\min_{w^T(x_c + x) - \gamma = 0} x^T \Sigma x \leq 1 \quad \Rightarrow \quad \frac{|\gamma - w^T x_c|}{\sqrt{w^T \Sigma^{-1} w}} \leq 1,$$

then the objective function attains zeros for all $x_* \in \mathcal{S}$. In particular, we can choose

$$x_* = \frac{|\gamma - w^T x_c|}{\sqrt{w^T \Sigma^{-1} w}} \cdot \frac{\operatorname{sgn}(\gamma - w^T x_c)}{\sqrt{w^T \Sigma^{-1} w}} \Sigma^{-1} w.$$

□

A.2 Robust LDA

We show that (41) is equivalent to (42). The formula of G in the numerator is by elementary analysis. The minimization problem in the denominator amounts to computing the shortest projection of $\mu_x - \mu_y$ onto z. Since $(\mu_x - \overline{\mu}_x)^T S_x^{-1}(\mu_x - \overline{\mu}_x) \leq 1$ is an ellipsoid, the projection of μ_x onto z satisfies

$$z^T\mu_x \in [z^T\overline{\mu}_x - \sqrt{z^TS_xz},\ z^T\overline{\mu}_x + \sqrt{z^TS_xz}] = [a_x, b_x].$$

Similarly, the projection of μ_y onto z satisfies

$$z^T\mu_y \in [z^T\overline{\mu}_y - \sqrt{z^TS_yz},\ z^T\overline{\mu}_y + \sqrt{z^TS_yz}] = [a_y, b_y].$$

Therefore, we can write the minimization problem equivalently as

$$\begin{array}{ll} \min & (z^T\mu_x - z^T\mu_y)^2 \\ \text{s.t.} & z^T\mu_x \in [a_x, b_x],\ z^T\mu_y \in [a_y, b_y]. \end{array}$$

The minimizer is 0 if the interval $[a_x, b_x]$ intersects $[a_y, b_y]$, i.e., $|z^T\overline{\mu}_x - z^T\overline{\mu}_x| \leq \sqrt{z^TS_xz} + \sqrt{z^TS_yz}$. Otherwise, the minimizer is given by the minimal distance between the end points of the intervals

$$\left(|z^T\overline{\mu}_x - z^T\overline{\mu}_x| - \left(\sqrt{z^TS_xz} + \sqrt{z^TS_yz}\right)\right)^2 = (f(z)^Tz)^2,$$

where the last equation is verified by direct calculation.

References

[1] P.-A. Absil and K. A. Gallivan. Accelerated line-search and trust-region methods. *SIAM J. Numer. Anal.*, 47(2):997–1018, 2009.

[2] S. Ahmed. *Robust estimation and sub-optimal predictive control for satellites.* PhD thesis, Imperial College London, 2012.

[3] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, editors. *Templates for the Solution of Algebraic Eigenvalue Problems: a Practical Guide.* SIAM, Philadelphia, 2000.

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

[5] A. Ben-Tal and A. Nemirovski. Robust solutions of linear programming problems contaminated with uncertain data. *Math. Progr.*, 88(3):411–424, 2000.

[6] P. Benner, A. Onwunta, and M. Stoll. An inexact Newton-Krylov method for stochastic eigenvalue problems. *eprint arXiv:1710.09470*, 2017.

[7] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Process. Mag.*, 25(1):41–56, 2008.

[8] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley & Sons, New York, 2012.

[9] G. Ghanem and D. Ghosh. Efficient characterization of the random eigenvalue problem in a polynomial chaos decomposition. *Int. J. Numer. Meth. Engng*, 72:486–504, 2007.

[10] G. H. Golub and C. F. Van Loan. *Matrix Computations.* Third Edition, Johns Hopkins University, Press, Baltimore, MD, USA, 1996.

[11] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, 2017. http://cvxr.com/cvx.

[12] E. Jarlebring, S. Kvaal, and W. Michiels. An inverse iteration method for eigenvalue problems with eigenvector nonlinearities. *SIAM J. Sci. Comput.*, 36(4):A1978–A2001, 2014.

[13] M. Kawanabe, W. Samek, K.-R. Müller, and C. Vidaurre. Robust common spatial filters with a maxmin approach. *Neural Computation*, 26(2):349–376, 2014.

[14] S.-J. Kim and S. Boyd. A minimax theorem with applications to machine learning, signal processing, and finance. *SIAM J. Optim.*, 19(3):1344–1367, 2008.

[15] S.-J. Kim, A. Magnani, and S. Boyd. Robust Fisher discriminant analysis. In *The Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 659–666, 2006.

[16] C. Le Bris. Computational chemistry from the perspective of numerical analysis. *Acta Numer.*, 14:363–444, 2005.

[17] J. Li and P. Stoica, editors. *Robust Adaptive Beamforming.* John Wiley & Sons, New York, 2005.

[18] M. Lichman. UCI machine learning repository, 2013. available at http://archive.ics.uci.edu/ml.

[19] X. Liu, X. Wang, Z. Wen, and Y. Yuan. On the convergence of the self-consistent field iteration in Kohn–Sham density functional theory. *SIAM J. Matrix Anal. Appl.*, 35(2):546–558, 2014.

[20] O. L. Mangasarian and E. W. Wild. Multisurface proximal support vector machine classification via generalized eigenvalues. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(12):1–6, 2005.

[21] R. M. Martin. *Electronic Structure: Basic Theory and Practical Methods.* Cambridge University Press, Cambridge, UK, 2004.

[22] J. Nocedal and S. Wright. *Numerical Optimization.* Springer Science & Business Media, Berlin, 2006.

[23] V. R. Saunders and I. H. Hillier. A level–shifting method for converging closed shell Hartree–Fock wave functions. *Int. J. Quantum Chem.*, 7(4):699–705, 1973.

[24] S. Shahbazpanahi, A. B. Gershman, Z.-Q. Luo, and K. M. Wong. Robust adaptive beamforming for general-rank signal models. *IEEE Trans. Signal Process.*, 51(9):2257–2269, 2003.

[25] N. D. Sidiropoulos, T.N. Davidson, and Z.-Q. Luo. Transmit beamforming for physical-layer multicasting. *IEEE Trans. Signal Process.*, 54(6):2239–2251, 2006.

[26] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proc Annu Symp Comput Appl Med Care*, pages 261–265. IEEE Computer Society Press, 1988.

[27] M. Soltanalian, A. Gharanjik, and M. R. B. Shankar. Grab-n-Pull: A Max-Min fractional quadratic programming framework with applications in signal processing. In *Proc IEEE Int Conf Acoust Speech Signal Process (ICASSP)*, page 1, 2015.

[28] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory.* Academic Press, Boston, 1990.

[29] L. Thøgersen, J. Olsen, D. Yeager, P. Jørgensen, P. Sałek, and T. Helgaker. The trust-region self-consistent field method: Towards a black-box optimization in Hartree–Fock and Kohn–Sham theories. *J. Chem. Phys.*, 121(1):16–27, 2004.

[30] S.A. Vorobyov, A.B. Gershman, and Z.-Q. Luo. Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem. *IEEE Trans. Signal Process.*, 51(2):313–324, 2003.

[31] P. Xanthopoulos, M. R. Guarracino, and P. M. Pardalos. Robust generalized eigenvalue classifier with ellipsoidal uncertainty. *Ann. of Opera. Res.*, 216(1):327–342, 2014.

[32] C. Yang, W. Gao, and J. C. Meza. On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 30(4):1773–1788, 2009.

[33] C. Yang, J. C. Meza, and L.-W. Wang. A trust region direct constrained minimization algorithm for the Kohn-Sham equation. *SIAM J. Sci. Comput.*, 29(5):1854–1875, 2007.

[34] S. Yu, L. Tranchevent, B. De Moor, and Y. Moreau. *Kernel-based Data Fusion for Machine Learning.* Springer, Berlin, 2011.

[35] L.-H. Zhang. On a self-consistent-field-like iteration for maximizing the sum of the Rayleigh quotients. *J. Comput. Appl. Math.*, 257:14–28, 2014.

[36] L.-H. Zhang and R.-C. Li. Maximization of the sum of the trace ratio on the Stiefel manifold, I: Theory. *SCIENCE CHINA Math*, 57(12):2495–2508, 2014.