ON THE LEAST SQUARES APPROXIMATION OF SYMMETRIC-DEFINITE PENCILS SUBJECT TO GENERALIZED SPECTRAL CONSTRAINTS

MOODY T. CHU* AND QUANLIN GUO†

Abstract. A general framework for the least squares approximation of symmetric-definite pencils subject to generalized eigenvalues constraints is developed in this paper. This approach can be adapted to different applications, including the inverse eigenvalue problem. The idea is based on the observation that a natural parameterization for the set of symmetric-definite pencils with the same generalized eigenvalues is readily available. In terms of these parameters, descent flows on the isospectral surface aimed at reducing the distance to matrices of the desired structure can be derived. These flows can be designed to carry certain other interesting properties and may be integrated numerically.

Key words. Matrix Pencil, Generalized Eigenvalue, Symmetric-definite Pencil, Inverse Problem, Least Squares, Descent Method, Isospectral Surface.

AMS(MOS) subject classifications. 65F15, 15A04, 65K10, 49D07.

1. Introduction. Let A and B be two square matrices of size n. A matrix pencil of A and B is a family of matrices $A - \lambda B$, parameterized by $\lambda \in \mathbb{C}$. Elements in the set $\sigma(A, B)$ defined by

$$\sigma(A, B) := \{z \in \mathbb{C} | \det(A - zB) = 0\} \tag{1}$$

are called the *generalized eigenvalues* of the pencil. It is easy to see that there are n generalized eigenvalues if and only if $\text{rank}(B) = n$. If B is rank deficient, then $\sigma(A, B)$ may be finite, empty, or infinite. Generalized eigenvalues are preserved under equivalence transformations, i.e., $\sigma(A, B) = \sigma(Y^H AX, Y^H BX)$, provided X and Y are nonsingular matrices and Y^H denotes the conjugate transpose of Y.

In this paper we shall limit our discussion to $\mathbb{R}^{n\times n}$, the Euclidean space of all $n \times n$ real-valued matrices equipped with the Frobenius inner product

$$\langle X, Y \rangle := \sum_{i,j} x_{ij} y_{ij}. \tag{2}$$

For convenience, we also introduce the notation $G(n)$ and $s(n)$ representing, respectively, the general linear group of all nonsingular matrices and the linear subspace of all symmetric matrices in $\mathbb{R}^{n\times n}$. It is frequently the case in practice, and will be assumed henceforth, that A is symmetric and B is symmetric and positive definite. Pencils of this variety are referred to as *symmetric-definite pencils* [7]. For convenience, the corresponding pair of matrices are referred to a *symmetric-definite pair*.

* Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205 (**chu@math.ncsu.edu**). This research was supported in part by the National Science Foundation under grant DMS-9422280.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205.

Obviously $A-\lambda B$ is symmetric-definite if and only if $P^TAP-\lambda P^TBP$ is symmetric-definite for all $P \in G(n)$. This congruence transformation naturally delineates a "parameterization" for the set

$$\mathcal{M}(A,B) := \{(P^TAP, P^TBP) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} | P \in G(n)\}. \tag{3}$$

We shall show that $\mathcal{M}(A,B)$, consisting of all symmetric-definite pairs with the same generalized eigenvalues $\sigma(A,B)$, is made of smooth submanifolds in $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$.

This paper concerns the construction of a symmetric-definite pencil satisfying simultaneously conditions on its structure and spectrum. We cast the problem as a task of finding the shortest distance between the set of structured matrices and the *isospectral* set $\mathcal{M}(A,B)$, where $\sigma(A,B)$ is the prescribed spectrum. The approximation is measured by the Frobenius norm over the product space $s(n) \times s(n)$, so a solution is best in the sense of least squares.

More specifically, let V_i, $i = 1, 2$, denote either a single matrix or an affine subspace in $s(n)$ whose elements, qualified by satisfying certain specified conditions on their structure, are being approximated. Define $\mathcal{P} : s(n) \times s(n) \longrightarrow V_1 \times V_2$ by

$$\mathcal{P}(X,Y) := (\mathcal{P}_1(X), \mathcal{P}_2(Y)) \tag{4}$$

where $\mathcal{P}_1$ and $\mathcal{P}_2$ denote, respectively, the projections from $s(n)$ onto V_1 and V_2 with respect to the inner product (2) . In case V_i is a singleton, define $\mathcal{P}_i(X) \equiv V_i$. The approximation is considered through the optimization problem:

$$\min_{(X,Y)\in\mathcal{M}(A,B)} \frac{1}{2}\|(X,Y) - \mathcal{P}(X,Y)\|^2, \tag{5}$$

i.e., the part of (X,Y) that does not carry the desirable structure is being minimized. We emphasize here that the desirable structure in V_1 can be defined independently of that in V_2.

One important point should be clarified before we move on to the discussion of solving (5). We mention that there are two constraints, the spectrum and the structure, imposed upon an ideal problem. In practice, it may occur that one of the two constraints should be more critical than the other due to, for example, the physical realizability. On the other hand, there are also situations where one constraint could be more relaxed than the other due to, for example, the physical uncertainty. Structural constraint usually is imposed due to the physical realizability. Spectral constraint often carry some physical uncertainty. In reality, it is often difficult to maintain both the spectral constraint and the structural constraint concurrently. When these constraints cannot be satisfied simultaneously, a least squares solution becomes the next best thing we can hope for. Depending upon which constraint is to be enforced explicitly, we would have different ways of defining a least squares approximation. The situation in (5) is such that while the pair of matrices (X,Y) vary among the isospectral surface $\mathcal{M}(A,B)$ and hence keep the spectrum $\sigma(A,B)$, the discrepancy between (X,Y) and the desirable structure is minimized. Another situation, which is not addressed in this paper, is to

seek for a symmetric-definite pair of matrices (X, Y) in the space $V_1 \times V_2$ (and hence the structure is maintained) so that the discrepancy between the two sets $\sigma(X, Y)$ and $\sigma(A, B)$ is minimized. At the first glance, these two situations appear to be quite different. In particular, a parameterization for symmetric-definite pairs of matrices with structure specified by V_1 and V_2 is difficult, if not impossible, to obtain. However, it is remarkable that in certain special circumstances these two seemingly unrelated problems can be shown to be equivalent. One such a case is the inverse ordinary eigenvalue problem that has already been discussed in [2]. In this paper, we shall focus on (5) only.

The choices of V_i in the set-up make the problem (5) quite versatile in application. We mention three immediate applications below. We shall come back in a later part of this paper to explain more specifically how these problems can be solved by our technique.

PROBLEM 1. *Given a symmetric-definite pair of matrices $(\tilde{A}, \tilde{B})$ and real numbers $\lambda_1, \ldots, \lambda_n$, find the least squares approximation (X, Y) to $(\tilde{A}, \tilde{B})$ such that (X, Y) is still symmetric-definite but $\sigma(X, Y) = \{\lambda_1, \ldots, \lambda_n\}$.*

A question that resembles Problem 1 but in the context of ordinary eigenvalue problems, i.e., when $Y \equiv \tilde{B} = I$, can be answered by the Wielandt-Hoffman theorem [4, 10]. For generalized eigenvalue problems, however, the perturbation theory is much more complicated. See, for example, [17, Chapter VI, Section 3]. Our approach, by taking $A = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$ and $B = I$ in the definition of the isospectral surface $\mathcal{M}(A, B)$, and $V_1 \equiv \tilde{A}$ and $V_2 \equiv \tilde{B}$ in the definition of the projection $\mathcal{P}$, offers an interesting and easy way to solve Problem 1.

PROBLEM 2. *Given a symmetric-definite pencil $A - \lambda B$, find all its generalized eigenvalues.*

Among the well-known numerical methods for the symmetric (ordinary) eigenvalue problem, one idea of Jacobi is to systematically reduce the norm of off-diagonal elements. A similar idea can be applied to Problem 2 if we take V_1 and V_2 to be the subspace of all diagonal matrices. In this way, the minimization in (5) amounts to reducing the off-diagonal elements of both X and Y simultaneously by congruence transformation. We shall see that a simple analysis on the stationary points of (5) re-establishes the well-known fact that any symmetric-definite pair can be simultaneously diagonalized.

PROBLEM 3. *Given a symmetric-definite pair $(\tilde{A}, \tilde{B})$ and values $\lambda_1, \ldots, \lambda_n$, find a diagonal matrix D so that $\sigma(\tilde{A} + D, \tilde{B}) = \{\lambda_1, \ldots, \lambda_n\}$.*

Generalized eigenvalue problems arise, for example, when a Sturm-Liouville problem is discretized by high-order implicit finite difference schemes [14]. An inverse problem, such as Problem 3, is then to reconstruct a certain physical parameter from the natural frequencies. Research on inverse (ordinary) eigenvalue problems has been extensive and fruitful. See, for example, [8] and the references contained therein. Obviously, if $\tilde{B} = \tilde{L}\tilde{L}^T$ is the Cholesky decomposition of $\tilde{B}$, then Problem 3 can be reformulated as finding D such that $\sigma(\tilde{L}^{-1}(\tilde{A} + D)\tilde{L}^{-T}, I) = \{\lambda_1, \ldots, \lambda_n\}$, which becomes an inverse ordinary eigenvalue problem. On the other hand, we may choose, among several options to be discussed in the sequel, V_1 to be the affine subspace of $\tilde{A}$ plus all diagonal

matrices, $V_2 \equiv \tilde{B}$, $A = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$, and $B = I$. Our approach avoids the inversion of any matrix and guarantees a least squares solution even if an exact solution does not exist.

The multiplicative inverse eigenvalue problem is another important class of problem in applications. The question centers around finding a diagonal matrix D^{-1} so that the "preconditioned" matrix $D^{-1}M$ possesses a specialized spectrum. A multiplicative inverse eigenvalue problem can be formulated as an inverse generalized eigenvalue problem $M - \lambda D$ in a setting similar to Problem 3 except the first entry M is held constant instead.

Solving (5) by standard techniques for constrained optimization problems is not easy because of the matrix structure involved. The main point of this paper is to cultivate descent flows on $\mathcal{M}(A, B)$ for solving (5) in general. Our approach offers a new channel for tackling generalized spectrally constrained problems. The scheme of following flows in the open set $G(n)$ has a similar spirit of an interior-point method [9, 19], an area that has attracted enormous attention in recent years. However, our methods differ from the traditional interior-point methods in several aspects: Neither our objective function nor our feasible set is convex [1, 13, 18], and for the most part of our flows the dynamics is directed by the objective value rather than the penalty function [20]. We shall comment on this connection again at the end of Example 1 in §5.

This paper is organized as follows: We begin in §2 to study the geometry of the isospectral set $\mathcal{M}(A, B)$. We shall show by the algebraic curve theory that $\mathcal{M}(A, B)$ is a union of smooth manifolds. We even can count its dimension in the generic case. In §3 we outline a framework from which specific differential equations can be designed based on needs or circumstances. The differential equations produce descent flows for (5). Our approach is flexible, yet it offers some theoretical insights as well as ready-made numerical algorithms. In an earlier paper [4], projected gradient flows were derived for least squares approximations with ordinary spectral constraints. Our development here is similar, except that no projection of the gradient is needed this time because $G(n)$ itself is an open set in $\mathbb{R}^{n\times n}$. On the other hand, it will become clear in our study that in order for a flow to maintain a certain additional property, such as being defined on $\mathcal{M}(A, B)$ without reference to its parameterization, the descent direction somehow has to be a modification of the gradient. This point will become manifest in §3. We highlight some specific applications in §4. Finally in §5 we report some numerical experiments.

2. Isospectral Surface. By flows we mean integral curves of a differential system. To define flows on the set $\mathcal{M}(A, B)$, we have to be certain first of all that $\mathcal{M}(A, B)$ is made of smooth entities. Toward this, we establish two results in this section concerning the topology of $\mathcal{M}(A, B)$.

THEOREM 2.1. *Given any symmetric-definite pair of matrices (A, B), the set $\mathcal{M}(A, B)$ consists of all symmetric-definite pairs with generalized eigenvalues $\sigma(A, B)$.*

Proof. It is clear that if $(X, Y) \in \mathcal{M}(A, B)$, then $X - \lambda Y$ is symmetric-definite and $\sigma(X, Y) = \sigma(A, B)$. It is known that any symmetric-definite pencil can be simultaneously diagonalized by congruence transformations. Therefore, if a symmetric-definite

pencil $X - \lambda Y$ has the same generalized spectrum $\sigma(A, B)$, then $X - \lambda Y$ is congruent to $\mathrm{diag}(\sigma(A, B)) - \lambda I$ and hence to $A - \lambda B$. This proves the assertion. □

The definition (3) may be thought of as an algebraic way to parameterize the set $\mathcal{M}(A, B)$. Note that the parameters come from $G(n)$ which is an open set in $\mathbb{R}^{n\times n}$. The parameterization implies, therefore, that $\mathcal{M}(A, B)$ can be a geometric entity of dimension *at most* n^2. More precisely, we have the following theorem.

THEOREM 2.2. *For any given symmetric-definite pair* (A, B), $\mathcal{M}(A, B)$ *is a disjoint union of smooth manifolds, each of which has only finitely many components and has dimension at most* n^2 *in* $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$.

Proof. Consider the vector $c(X, Y) := [c_1(X, Y), \ldots, c_n(X, Y)]^T$ whose components are defined by the coefficients in the polynomial

$$\det(X - zY) = (-1)^n \det(Y) z^n + \sum_{i=0}^{n-1} c_{n-i}(X, Y) z^i.$$

Clearly each $c_k(X, Y)$ is a polynomial in the entries of X and Y. Suppose $\sigma(A, B) = \{\lambda_1, \ldots, \lambda_n\}$. Consider the algebraic variety

$$\mathcal{V}(\lambda_1, \ldots, \lambda_n) := \{(X, Y) \in s(n) \times s(n) | c(X, Y) = (-1)^n \det(Y)\gamma\} \tag{6}$$

where $\gamma := [\gamma_1, \ldots, \gamma_n]^T$ with $\gamma_k := (-1)^k \sum_{i_1<\ldots<i_k} \lambda_{i_1} \ldots \lambda_{i_k}$. It follows from Whitney's stratification theorem [12, Theorem 2.3 and 2.4] that $\mathcal{V}(\lambda_1, \ldots, \lambda_n)$ can be expressed as a finite disjoint union of smooth manifolds, each of which has only finitely many components. Observe that

$$\mathcal{M}(A, B) = \mathcal{V}(\lambda_1, \ldots, \lambda_n) \cap (s(n) \times \mathcal{C}(n))$$

where $\mathcal{C}(n)$ is the cone of symmetric and positive definite matrices in $\mathbb{R}^{n\times n}$. Since $\mathcal{C}(n)$ obviously is a submanifold in $s(n)$, the assertion follows. □

The gauge n^2 of the dimension is not necessarily an overestimate. We can maintain a little bit more precision on the dimensions of submanifolds involved in Theorem 2.2. A somewhat related discussion can be found in [6]. Let

$$\rho := \max_{(X,Y)\in\mathcal{V}(\lambda_1,\ldots,\lambda_n)} \mathrm{rank} \left[\frac{\partial c}{\partial(X, Y)}\right].$$

Define

$$\mathcal{N}(\lambda_1, \ldots, \lambda_n) := \{(X, Y) \in \mathcal{V}(\lambda_1, \ldots, \lambda_n) | \mathrm{rank} \left[\frac{\partial c}{\partial(X, Y)}\right] < \rho\}. \tag{7}$$

Whitney's theorem affirms that $\mathcal{V}(\lambda_1, \ldots, \lambda_n) - \mathcal{N}(\lambda_1, \ldots, \lambda_n)$ is a smooth manifold of dimension $n(n+1)-\rho$. Furthermore, because the rank deficient condition in (7) imposes extra polynomial equations on (X, Y), the set $\mathcal{N}(\lambda_1, \ldots, \lambda_n)$ itself, if not empty, is a union of manifolds with lower dimensions. It follows that $\mathcal{V}(\lambda_1, \ldots, \lambda_n) - \mathcal{N}(\lambda_1, \ldots, \lambda_n)$ is the largest manifold component of $\mathcal{V}(\lambda_1, \ldots, \lambda_n)$ in the sense that $\mathcal{N}(\lambda_1, \ldots, \lambda_n)$ is

nowhere dense and has measure zero relative to $\mathcal{V}(\lambda_1, \ldots, \lambda_n)$. Observe that involved in (6) are $n(n+1)$ unknowns and n equations, so it must be that $\rho < n$. It follows that the dimension of $\mathcal{V}(\lambda_1, \ldots, \lambda_n) - \mathcal{N}(\lambda_1, \ldots, \lambda_n)$ is *at least* n^2. Together with Theorem 2.2, we conclude that if

$$\mathcal{M}(A, B) \cap \mathcal{N}(\lambda_1, \ldots, \lambda_n) = \emptyset, \tag{8}$$

then $\mathcal{M}(A, B)$ is a smooth manifold of dimension exactly n^2. Sard's theorem [11] guarantees that for almost all choices of (A, B), the condition (8) holds. In particular, it can be shown that (8) holds if (A, B) has distinct generalized eigenvalues. The above result on the parameterization and dimensionality for isospectral symmetric-definite pairs of matrices seems to be known the first time. Though the result may not appear too surprising, the way it is obtained by utilizing the Whitney's theorem is of interest in its own right.

We stress before we move on to describe flows on $\mathcal{M}(A, B)$ that for our application it is not essential whether the set $\mathcal{M}(A, B)$ itself is a one-piece manifold. The differentiable flows that will be defined later automatically stay on smooth components of $\mathcal{M}(A, B)$.

We conclude this section by one example showing that the inverse eigenvalue problems for matrix pencils could be quite intricate. We show that in special circumstances $\mathcal{M}(A, B)$ may be a proper subset of $\mathcal{N}(\lambda_1, \ldots, \lambda_n)$. Consider the case when $n = 2$, $A = 0$ and $B = I$. Then $\mathcal{M}(A, B) = \{(0, P^T P) | P \in G(n)\}$. Though $G(2)$ has dimension 4, $\mathcal{M}(A, B)$ obviously has dimension 3. It is interesting to note that for a pair $X = (x_{ij})$ and $Y = (y_{ij})$ to be in $\mathcal{V}(0, 0)$ a necessary condition is that the entries satisfy the equations:

$$\begin{aligned} x_{21} &= x_{12}, \\ y_{21} &= y_{12}, \\ x_{11} &= |x_{12}| \frac{\mathrm{sgn}(x_{12}) y_{12} \pm \sqrt{y_{12}^2 - y_{11} y_{22}}}{y_{11}}, \\ x_{22} &= \frac{-y_{11} x_{11} + 2 x_{12} y_{12}}{y_{22}}, \end{aligned}$$

provided $y_{11} y_{22} \neq 0$. There are four free parameters in defining $\mathcal{V}(0, 0)$. However, if Y is required to be positive definite, then $X = 0$ is the only possible solution.

3. Descent Flows. The parameterization (3) provides grounds for maneuver on $\mathcal{M}(A, B)$ to reduce the objective value in (5). In this section, we discuss how to take advantage of this parameterization to formulate descent flows.

We start with working within the parameter space $G(n)$. For convenience, we introduce the abbreviation:

$$\begin{cases} \alpha_1(P) := P^T A P - \mathcal{P}_1(P^T A P) \\ \alpha_2(P) := P^T B P - \mathcal{P}_2(P^T B P) \end{cases}, \tag{9}$$

when the symmetric-definite pair (A, B) is fixed. The objective function in (5) is equivalent to the function $F : G(n) \longrightarrow R$ where

$$F(P) := \frac{1}{2}\left(\langle \alpha_1(P), \alpha_1(P)\rangle + \langle \alpha_2(P), \alpha_2(P)\rangle\right). \tag{10}$$

The following result is critical in our development.

THEOREM 3.1. *The gradient ∇F of F is given by*

$$\nabla F(P) = 2\left\{AP\alpha_1(P) + BP\alpha_2(P)\right\}. \tag{11}$$

Proof. Observe that the Fréchet derivative of F at P acting on $H \in \mathbb{R}^{n\times n}$ can be calculated as follows:

$$\begin{aligned}
F'(P)H &= \langle \alpha_1(P), H^TAP - \mathcal{P}_1'(P^TAP)H^TAP + P^TAH - \mathcal{P}_1'(P^TAP)P^TAH\rangle \\
&\quad + \langle \alpha_2(P), H^TBP - \mathcal{P}_2'(P^TBP)H^TBP + P^TBH - \mathcal{P}_2'(P^TBP)P^TBH\rangle \\
&= 2\Big\{\langle \alpha_1(P), P^TAH - \mathcal{P}_1'(P^TAP)P^TAH\rangle \\
&\quad + \langle \alpha_2(P), P^TBH - \mathcal{P}_2'(P^TBP)P^TBH\rangle\Big\} \\
&= 2\Big\{\langle \alpha_1(P), P^TAH\rangle + \langle \alpha_2(P), P^TBH\rangle\Big\} \\
&= 2\langle AP\alpha_1(P) + BP\alpha_2(P), H\rangle. \qquad (12)
\end{aligned}$$

In the above, the second equality is due to the symmetry of the matrices involved. The third equality follows from the fact that the action of $\mathcal{P}_i'$ (at P^TAP and P^TBP, respectively) on any point (P^TAH and P^TBH, specifically) resides in the tangent space of V_i whereas the range of α_i is perpendicular to the tangent space of V_i. The last equality is obtained by utilizing the adjoint property of the Frobenius inner product. It follows from (12) that the gradient ∇F of F may be interpreted as asserted. □

Obviously, the differential equation

$$\dot{P}(t) := -\nabla F(P(t)), \tag{13}$$

where $\dot{P}$ means the derivative of P with respect to a certain artificial parameter t, defines the steepest descent flow $P(t)$ on $G(n)$ for F. It should be cautioned however that the open set $G(n)$ has *boundary* made of all $n\times n$ singular matrices. The differential equation (13) alone cannot guarantee that the flow $P(t)$ will stay away from the boundary of singular matrices. The first example in §5 clearly illustrates this occurrence.

Through the parameterization relationship

$$\left\{\begin{array}{rcl} X(t) &=& P(t)^TAP(t) \\ Y(t) &=& P(t)^TBP(t), \end{array}\right. \tag{14}$$

each flow in the parameter space $G(n)$ has a corresponding flow on $\mathcal{M}(A, B)$. Related to the flow $P(t)$ defined by (13), for example, is the flow $X(t)$ defined by :

$$\dot{X} = -2\Big\{\alpha_1(P)P^TA^2P + \alpha_2(P)P^TBAP$$

$$+P^TA^2P\alpha_1(P) + P^TABP\alpha_2(P)\Big\} \tag{15}$$

$$= -2\Big\{\beta_1(X)X(P^TP)^{-1}X + \beta_2(Y)Y(P^TP)^{-1}X$$

$$+X(P^TP)^{-1}X\beta_1(X) + X(P^TP)^{-1}Y\beta_2(Y)\Big\}, \tag{16}$$

where we have denoted

$$\begin{cases} \beta_1(X) := \alpha_1(P) \\ \beta_2(Y) := \alpha_2(P) \end{cases}, \tag{17}$$

to emphasize the dependence of the system on the variables X and Y. A similar flow $Y(t)$ can also be defined.

Neither (15) nor (16) is useful in that the differential system depends explicitly on the parameterization variable P. That dependence means that to integrate (15) or (16) one must also integrate (13). This is a waste since the parameter flow $P(t)$ needs to be integrated in any case. It perhaps would be more economical to obtain $X(t)$ and $Y(t)$ directly from (14).

Note also that the system (13) defines the steepest descent flow. There are situations when one prefers to relinquish the steepest descent property in exchange for maintaining other attributes. In the following we introduce several other descent flows for this purpose.

We first illustrate a situation where the description of $X(t)$ and $Y(t)$ can be implicit in the parameter P.

COROLLARY 3.2. *The flow defined by*

$$\dot{P} := -\frac{1}{2}PP^T\nabla F(P). \tag{18}$$

is a descent flow.

Proof. Observe that

$$\langle \nabla F(P), -PP^T\nabla F(P)\rangle = -\langle P^T\nabla F(P), P^T\nabla F(P)\rangle \leq 0$$

and that the equality holds only when $\nabla F(P) = 0$. Thus, the differential system (18), though not the steepest one, continues to define a descent flow for F. □

Upon substitution, the corresponding flow $(X(t), Y(t))$ on $\mathcal{M}(A, B)$ is defined by the differential system:

$$\begin{cases} \dot{X} = -\Big((XW)^T + XW\Big) \\ \dot{Y} = -\Big((YW)^T + YW\Big) \end{cases} \tag{19}$$

with

$$W := X\beta_1(X) + Y\beta_2(Y). \tag{20}$$

Note that the differential system (19) is autonomous in X and Y, and makes no reference to the variable P. The computation of $P(t)$ as well the troublesome matrix inversion such as $(P^TP)^{-1}$ in (16) are thus avoided.

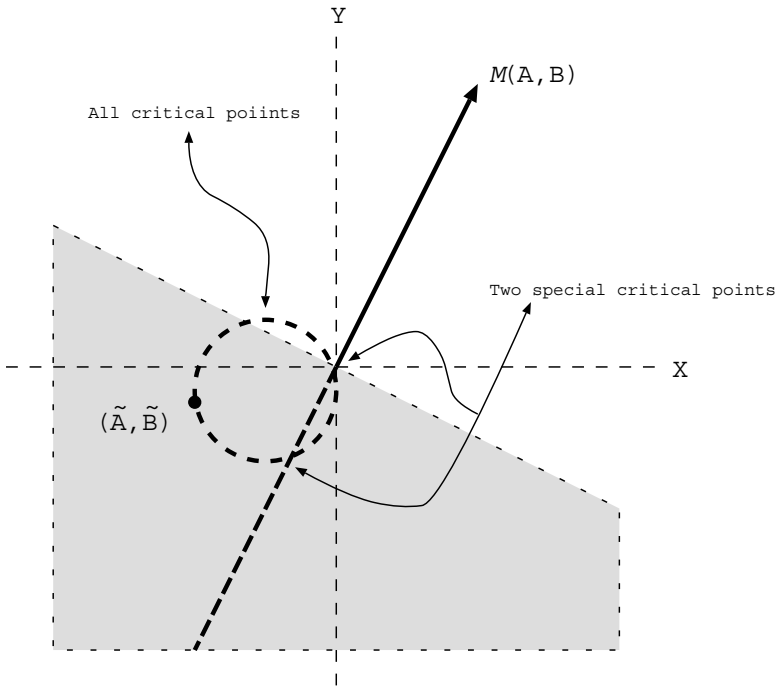

FIG. 1. *Geometry of a pseudo solution.*

It is worth noting that the critical points of the differential system (18) are exactly the same as the stationary points of the optimization problem (5), provided that critical point is nonsingular. The optimization problem (5), therefore, can be solved by integrating (19) from a suitable starting point, say $(X(0), Y(0)) = (A, B)$, until a limit point is located.

The simplest case of (19) when $n = 1$ is rather illuminating. Corresponding to a given pair of numbers (A, B) with $B > 0$, the set $\mathcal{M}(A, B) = \{(X, Y) \in \mathbb{R}^2 | X = AP^2, Y = BP^2, P \neq 0\}$ is an half array that emanates from but does not include the origin in the direction (A, B). In particular, $\mathcal{M}(A, B)$ is an *unbounded open* set. Suppose we want to solve Problem 1 mentioned in §1. The corresponding differential system of (19) becomes

$$\left\{ \begin{array}{rcl} \dot{X} & = & -2X(X(X-\tilde{A}) + Y(Y-\tilde{B})) \\ \dot{Y} & = & -2Y(X(X-\tilde{A}) + Y(Y-\tilde{B})). \end{array} \right. \tag{21}$$

All critical points of (21) are included in the set

$$\{(X, Y) | X(X-\tilde{A}) + Y(Y-\tilde{B}) = 0\},$$

which is the dotted circle represented in Figure 1. But relative to $\mathcal{M}(A, B)$ where the flow starting from $X(0) = A$ and $Y(0) = B$ resides, only the two critical points,

$$(0, 0) \text{ and } \left(\frac{C(C\tilde{A} + \tilde{B})}{I + C^2}, \frac{C\tilde{A} + \tilde{B}}{I + C^2} \right) \tag{22}$$

with $C := AB^{-1}$, are most pertinent. Consider the case when the target point $(\tilde{A}, \tilde{B})$ is located in the *lower* half plane of the line that passes through the origin and is perpendicular to the array $\mathcal{M}(A, B)$ (see the shaded region in Figure 1.) Obviously the shortest distance from $(\tilde{A}, \tilde{B})$ to $\mathcal{M}(A, B)$ is attained only at the origin, but that point does not belong to $\mathcal{M}(A, B)$. Thus, Problem 1 should have *no* true solution is this case. Nonetheless, the flow defined by (21) stays on the half array and indeed moves toward the origin. In this way, we end up with a *pseudo* solution in the sense that the solution is still a least squares approximation but that point is not from within $\mathcal{M}(A, B)$. On the other hand, the second critical point (22) in this case is away from the set $\mathcal{M}(A, B)$ by a positive distance and hence can never be realized. We shall refer back to (21) in §4 for further discussion of higher dimension case.

We next mention two more descent flows that possess some additional interesting properties.

COROLLARY 3.3. *The differential equation*

$$\dot{P} := -\frac{1}{2} P \left\{ P^T \nabla F(P) - \nabla F(P)^T P \right\} \tag{23}$$

defines a descent flow. Furthermore,

$$P(t)P(t)^T \equiv \textit{constant}.$$

Proof. From the fact that

$$\langle \nabla F(P), P \left\{ P^T \nabla F(P) - \nabla F(P)^T P \right\} \rangle = \langle P^T \nabla F(P), P^T \nabla F(P) - \nabla F(P)^T P \rangle$$

and the equality that

$$\langle M, M - M^T \rangle = \sum_{j \neq i} (m_{ij} - m_{ji})^2 \geq 0$$

for any square matrix $M = (m_{ij})$, it follows that the flow $P(t)$ enjoys the descent property. Furthermore, because the quantity in the braces of (23) is skew-symmetric, it is easy to see that $P\dot{P}^T + \dot{P}P^T = 0$. Thus $P(t)P(t)^T \equiv P(0)P(0)^T$ for all t. □

The corresponding flow on $\mathcal{M}(A, B)$ are integral curves of the double-bracket system:

$$\left\{ \begin{array}{rcl} \dot{X} & = & [X, [X, \mathcal{P}_1(X)] + [Y, \mathcal{P}_2(Y)]] \\ \dot{Y} & = & [Y, [X, \mathcal{P}_1(X)] + [Y, \mathcal{P}_2(Y)]] \end{array} \right. \tag{24}$$

where $[X, Y] := XY - YX$ denotes the Lie bracket. Note that the system (24) is autonomous. Note also that if $P(0) = I$ from the beginning, then $P(t)$ remains orthogonal for all t. Our notion here generalizes that of orthogonal similarity transformation discussed in [5].

COROLLARY 3.4. *The differential equation*

$$\dot{P} := -\frac{1}{2} \left\{ \nabla F(P) P^T - P \nabla F(P)^T \right\} P \tag{25}$$

is a descent flow. Furthermore,

$$P(t)^T P(t) \equiv \textit{constant}.$$

Proof. The proof is similar to Corollary 3.3. ∎

Although it looks similar to (23), this new system (25) by no means is a trivial alternation (say, by taking the transpose) of (23). In particular, it can be checked by substitution that the corresponding differential equation for $X(t)$ and $Y(t)$ depends explicitly on the variable P in (25), a predicament that does not occur in (24). The system (25) is especially useful for attacking problems where the corresponding flow $Y(t)$ is expected to be constant. Problem 3 is one such instance. We shall be more specific on its application in the next section.

We conclude this section with one remark on the asymptotic behavior of the flows.

THEOREM 3.5. *For all the flows $P(t)$ defined above, the corresponding $(X(t), Y(t))$ converges. Generically, the limit point is a stationary point, possibly on the boundary of $\mathcal{M}(A, B)$, of (5). The non-generic exception is when the product $P^T \nabla F(P)$ in (23) or $\nabla F(P) P^T$ in (25) is symmetric at the limit point.*

Proof. Along any solution $(X(t), Y(t))$ the function

$$G(t) := F(P(t)) = \frac{1}{2} \{ \langle \beta_1(X(t)), \beta_1(X(t)) \rangle + \langle \beta_2(Y(t)), \beta_2(Y(t)) \rangle \} \tag{26}$$

satisfies

$$\dot{G}(t) = \langle \nabla F(P(t)), \dot{P}(t) \rangle \leq 0.$$

Furthermore, $\dot{G} = 0$ only when $\nabla F(P) = 0$ or $\dot{P} = 0$. The latter case generically implies also $\nabla F(P) = 0$. Thus $G(t)$ is monotonically decreasing until a stationary point of (5) is found. ∎

4. Applications. Our differential system approach not only can be used as a convenient algorithm for finding a least squares solution, but also offers some theoretical insights into the problem. In this section we explain more specifically how our approach can be applied to solve the three problems described in §1. We discuss the applications case by case. Further numerical experiment will be reported in §5.

APPLICATION 1. We point out earlier that there is no easy generalization of the Wielandt-Hoffman theorem for Problem 1. To demonstrate the complexity of Problem 1 in general, we consider a very special case when both target matrices $\tilde{A}$ and $\tilde{B}$ are diagonal. Our point by this overly simplified problem is to illustrate how complicated the stationary points for Problem 1 could be. Suppose that the differential equation (19) (which is based on the descent flow (18)) is used to solve the problem from the initial values $X(0) = A = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$ and $Y(0) = B = I$. Recall that the critical points of the differential system are exactly the same as the stationary points of the problem. By construction we know the solution flow $(X(t), Y(t))$ of (19) remains diagonal. The differential system, being un-coupled into n pairs (x_{ii}, y_{ii}), $i = 1, \ldots n$,

can be represented exactly by (21) if all symbols there are interpreted as (diagonal) matrices. Observe that the pairs $(x_{ii}(t), y_{ii}(t))$ are independent of each other and may converge to limit points of different types (See (22)). In particular some of the pairs, as pointed out earlier, may converge to an infeasible limit point $(0, 0)$. This simple uncoupled system highlights the potential difficulty for general $\tilde{A}$ and $\tilde{B}$ where these events are intertwined together and hence make Problem 1 more complicated. Regardless of this complexity, our differential equation offers an easy-to-use numerical method for solving this type of problem.

APPLICATION 2. Using the set-up described in Problem 2, i.e., V_1 and V_2 are the subspaces of all diagonal matrices, the first-order optimality condition $\nabla F(P) = 0$ at any stationary point P is equivalent to the equality

$$X(X - \mathrm{diag}(X)) + Y(Y - \mathrm{diag}(Y)) = 0 \tag{27}$$

where X and Y are related to P by (14). It is easy to check that the diagonal elements involved in (27) are given by

$$\sum_{k \neq i} x_{ik}^2 + \sum_{k \neq i} y_{ik}^2 = 0, \ i = 1, \ldots, n. \tag{28}$$

That is, (X, Y) is a limit point of the descent flow (19) if and only if both X and Y are diagonal matrices. Our differential equation (19) not only re-establishes the fact that any symmetric-definite pencil can be simultaneously diagonalized, but also offered a numerical way to accomplish this.

APPLICATION 3. We give a little bit more details below for Problem 3 since it is of particular interest and importance. The geometry of Problem 3 is sketched in Figure 2 where we use the 3-D coordinate axes represent the triplet $(\text{off-diag}(X), \mathrm{diag}(X), Y)$ for any matrix pair $(X, Y) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n}$. The desirable state, represented by the bold horizontal line in Figure 2, means that $Y = \tilde{B}$ and $\text{off-diag}(X) = \text{off-diag}(\tilde{A})$. The minimization in (5) is equivalent to minimizing the distance between the two points **P** and **Q** in Figure 2 while **P** stays in $\mathcal{M}(A, B)$ (not drawn) and **Q** stays in the desirable state.

The desirable state can be characterized by selecting V_1 to be the affine subspace of $\tilde{A}$ plus all diagonal matrices and $V_2 \equiv \tilde{B}$. To maintain the eigenvalue information, an obvious choice would be letting $A = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$ and $B = I$. The projections corresponding to this set-up imply that $\beta_1(X) = \text{off-diag}(X - \tilde{A})$ and $\beta_2(Y) = Y - \tilde{B}$. While any of the differential equations we proposed, say (19), is ready for integration, there is a setback in using some of these equations — The resulting solution flow may *stop* at a local minimizer that does not meet the criteria of the desirable state, i.e., the resulting $Y(t)$ is likely to vary in t whereas the second matrix involved in Problem 3 is expected to be constantly $\tilde{B}$.

To remedy the above fault, we may consider using the differential system (25) with initial values

$$P(0) = UL^T \tag{29}$$

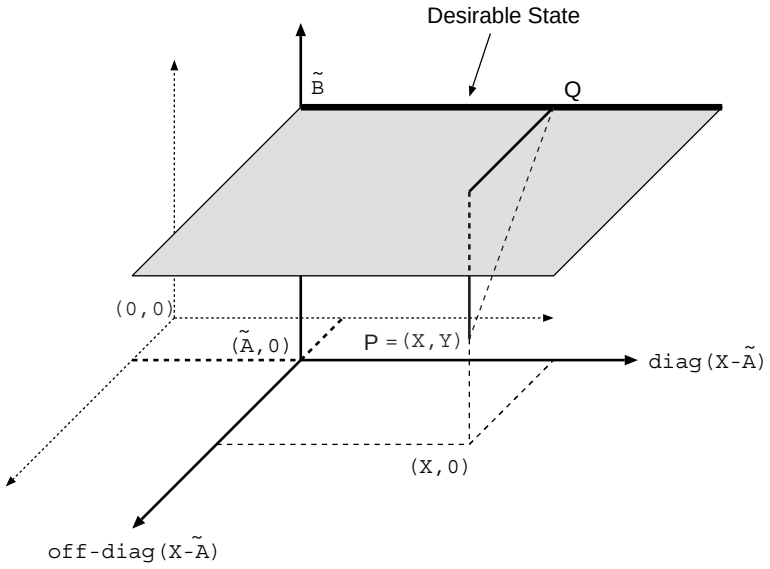

FIG. 2. *Geometry of Problem 3.*

where $\tilde{B} = LL^T$ is the Cholesky decomposition of $\tilde{B}$ and U is an arbitrary orthogonal matrix. Corollary 3.4 guarantees that $Y(t) \equiv \tilde{B}$ and hence $\beta_2(Y(t)) \equiv 0$ for all t. The differential equation (25) becomes

$$\dot{P} = \Big[P\alpha_1(P)P^T, \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}\Big]\, P \tag{30}$$

where $\alpha_1(P) = \text{off-diag}(P^T \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}P - \tilde{A})$. The Lie bracket operation in (30) is simple because the second operant is a diagonal matrix. The resulting $P(t)$ implicitly defines a flow $(X(t), Y(t))$ on the shaded region represented in Figure 2. The flow starts from $(X(0), Y(0)) = (L\mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}L^T, \tilde{B})$ and approximates the set of the desirable state. Once the limit point $P(\infty)$ of (30) is found, the diagonal matrix

$$D := \mathrm{diag}(X(\infty) - \tilde{A}),$$

where $X(\infty) = P(\infty)^T \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}P(\infty)$, is an optimal solution for Problem 3 in the sense of least squares.

5. Numerical Experiment. In this section we report some of our numerical experiments with the proposed methods. At present we are more concerned with the dynamics of the flows than the efficiency of the programs. Thus we only consider using general-purposed initial value problem software as the integrator. We have experimented with both the FORTRAN code ODE [15] and the MATLAB code ODE SUITE [16]. The results are similar. We shall only report experiments from ODE SUITE since it is easier to manipulate matrix operations and to present the results graphically by MATLAB.

There are two types of solvers, **ode113** and **ode15s**, in the MATLAB ODE SUITE. The code **ode113** is a PECE implementation of Adams-Bashforth-Moulton methods

for non-stiff systems. The code **ode15s** is a quasi-constant step size implementation of the Klopfenstein-Shampine family of the numerical differential formulas for stiff systems. The statistics about the cost of integration can be obtained directly from the **odeset** option built in the integrator. More details of these codes can be found in the document [16]. Again we have experimented with both solvers. We discover that when the prescribed eigenvalues do not vary wildly, these two codes perform comparably. But when the ratio of the eigenvalue with the largest magnitude to the smallest gets larger, the **ode15s** becomes faster in terms of CPU time. We think a largely varying spectrum perhaps has resulted in a stiff initial value problem.

In our experiments the tolerance for both absolute error and relative error is set at 10^{-12}. This criterion is used to control the accuracy in following the solution path. The high accuracy we required here has little to do with the dynamics of the underlying vector field, and perhaps is not needed in practical application. We examine the output values at time interval of 1 or 10, and assume that the path has reached an equilibrium point whenever the difference of the Lyapunov's functions (26) at two consecutive output points is less than 10^{-10}. So as to fit the data comfortably in the running text, we report only the case $n = 5$ and display all numbers with five digits.

Example 1. In our first experiment we report one pathological example where the flow $P(t)$ of parameters converges to the boundary of singular matrices, and hence the corresponding least squares problem is solved in an unusual yet interesting way.

Suppose we want to solve the generalized eigenvalue problem, Problem 2, for this pair of matrices

$$A = \begin{bmatrix} 1.0904 & 0.1575 & 0.2394 & 2.5284 & -0.4716 \\ 0.1575 & 0.2913 & -1.0421 & 1.8527 & 0.4591 \\ 0.2394 & -1.0421 & -2.2831 & -0.0859 & -2.2171 \\ 2.5284 & 1.8527 & -0.0859 & -2.5200 & -1.1272 \\ -0.4716 & 0.4591 & -2.2171 & -1.1272 & 1.1959 \end{bmatrix},$$

$$B = \begin{bmatrix} 6.8747 & -1.6174 & -1.3123 & 4.2938 & 0.5968 \\ -1.6174 & 6.8615 & 1.2753 & -2.2454 & -5.3684 \\ -1.3123 & 1.2753 & 2.8018 & 1.2469 & 0.6560 \\ 4.2938 & -2.2454 & 1.2469 & 5.1703 & 1.9403 \\ 0.5968 & -5.3684 & 0.6560 & 1.9403 & 10.6641 \end{bmatrix}$$

by using the steepest descent flow (13) with initial value

$$P(0) = \begin{bmatrix} -0.62735 & -0.04006 & 0.42746 & 0.63529 & 0.13607 \\ -0.41918 & -0.12833 & 0.34523 & -0.51495 & -0.65074 \\ -0.23520 & 0.77311 & -0.42324 & 0.18008 & -0.36799 \\ -0.22678 & 0.49212 & 0.28205 & -0.51204 & 0.60387 \\ -0.56918 & -0.37689 & -0.66288 & -0.19137 & 0.24073 \end{bmatrix}.$$

When our code terminates suggesting that a convergence has been reached, we discover

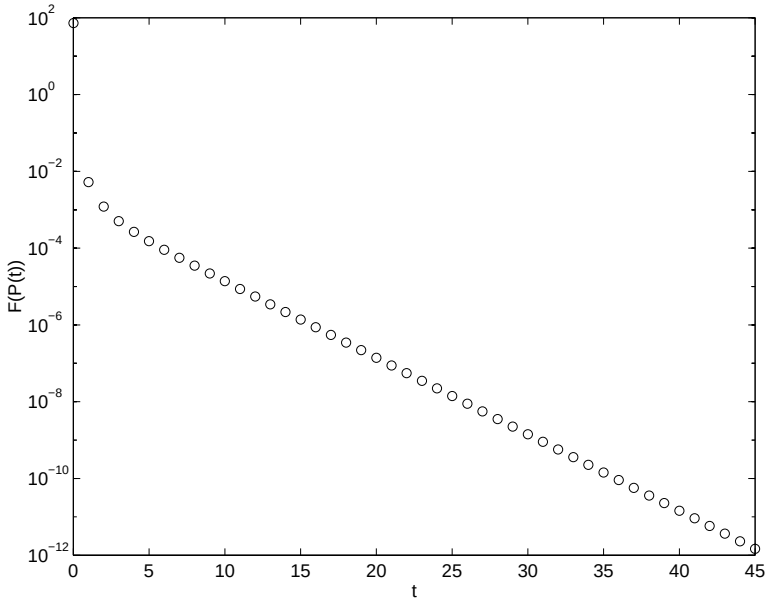

FIG. 3. *History of $F(P(t))$ Example 1 when $P(t)$ becomes singular.*

that

$$P(\infty) = \begin{bmatrix} -0.0243 & -0.1109 & 0.2316 & -0.0000 & 0.1459 \\ -0.1314 & 0.0106 & 0.1922 & -0.0000 & -0.3265 \\ 0.0860 & 0.2712 & 0.0432 & 0.0000 & -0.2058 \\ -0.0279 & 0.2026 & 0.3343 & -0.0000 & 0.2038 \\ -0.0979 & 0.0565 & -0.2861 & -0.0000 & 0.0261 \end{bmatrix}.$$

The fourth column of $P(\infty)$ is in fact as small as

$$\left[-0.20072\times10^{-13}, -0.13475\times10^{-12}, 0.94951\times10^{-13}, -0.32969\times10^{-13}, -0.97683\times10^{-13}\right]^T,$$

indicating that $P(\infty)$ is nearly singular. Note that this result of near singularity does not contradict with the condition (28) where we argue that (X, Y) is a stationary point of (5) if and only if both X and Y are diagonal matrices. Indeed, we obtain that

$$\begin{aligned} X = P(\infty)^T A P(\infty) &= \text{diag}\{0.0800, -0.4773, 0.7925, 0.0000, -0.4043\}, \\ Y = P(\infty)^T B P(\infty) &= \text{diag}\{0.0635, 0.6128, 2.4657, 0.0000, 2.1823\}. \end{aligned}$$

We can see also from Figure 3 that this limit point $P(\infty)$ is reducing the objective function (10) to zero. This limit point would be a global minimizer were it not becoming singular. The significant difference here is that since $P(\infty)$ is singular, the corresponding limit point (X, Y) is no longer congruently equivalent to (A, B). In particular, Y is now only positive semi-definite and hence the information of generalized eigenvalues is lost.

Results like this might be disappointing, but is still of some theoretic value. It illustrates how congruence transformation in reducing the off-diagonal elements of matrices

can go wrong. Our method may be far away from being practical per se among the many other ways to solve the generalized eigenvalue problem. But readers are reminded that the above illustration of solving Problem 2 by (13) is just one application of our general approach.

It is worthy to remark on three possible remedies along our notion above:

1. The QZ flow [3] is another differential equation approach that is analogous to the steepest descent flow described in this paper. The QZ flow, using orthogonal equivalence transformations instead, does not suffer from the fault of becoming singularity. The symmetric-definiteness, however, is not maintained.
2. Even with the descent flow approach, the singularity could be avoided by changing the initial value $P(0)$ and hence taking another path (and there are indeed infinitely many such initial guesses.) One could also use flows defined by (23) or (25) to carry out the computation, but we hasten to point out that because either $P(t)P(t)^T$ or $P(t)^TP(t)$ is constant for all t in these cases, not all symmetric-definite pairs (A, B) can be simultaneously diagonalized in this way.
3. Finally, it is possible to avoid the singularity by imposing penalties for singularity in the objective function (10) like those done in [1, 9, 19, 20] to avoid the semi-definiteness. This approach will eventually lead to the so called interior point methods that have been studied and developed extensively.

Example 2. In general, an inverse eigenvalue problem like Problem 3 can hardly have an exact solution at all. So an approximate solution in the sense of least squares is sometimes desirable. In this case the globally convergent flow defined by (30) becomes particularly meaningful. The flow approach guarantees convergence to a local solution.

To illustrate how the dynamical system (30) behaves, we first generate test data by considering a randomly generated symmetric-definite pair $(\hat{A}, \tilde{B})$:

$$\hat{A} = \begin{bmatrix} -2.8645 & 1.8576 & -2.1532 & 0.6710 & 0.5092 \\ 1.8576 & -0.1855 & 0.5149 & 2.1096 & -1.3318 \\ -2.1532 & 0.5149 & 1.3880 & -0.4591 & 0.3603 \\ 0.6710 & 2.1096 & -0.4591 & -4.3183 & -1.2334 \\ 0.5092 & -1.3318 & 0.3603 & -1.2334 & -1.8954 \end{bmatrix},$$

$$\tilde{B} = \begin{bmatrix} 6.0810 & -2.6691 & 0.6390 & -0.5509 & -1.0124 \\ -2.6691 & 5.5185 & 1.1005 & 0.8248 & 0.8014 \\ 0.6390 & 1.1005 & 2.4625 & 1.9543 & -0.4839 \\ -0.5509 & 0.8248 & 1.9543 & 4.2586 & -0.0535 \\ -1.0124 & 0.8014 & -0.4839 & -0.0535 & 0.8230 \end{bmatrix}.$$

We use its generalized eigenvalues

$$\sigma(\hat{A}, \tilde{B}) = \{3.9955, 0.3093, -0.6662, -1.2920, -3.2878\}$$

as the target spectrum in our experiment. We use $\tilde{A} = \hat{A} - \mathrm{diag}(\hat{A})$ and $\tilde{B}$ as the test data for Problem 3. Apparently, $\mathrm{diag}(\hat{A})$ is one global solution.

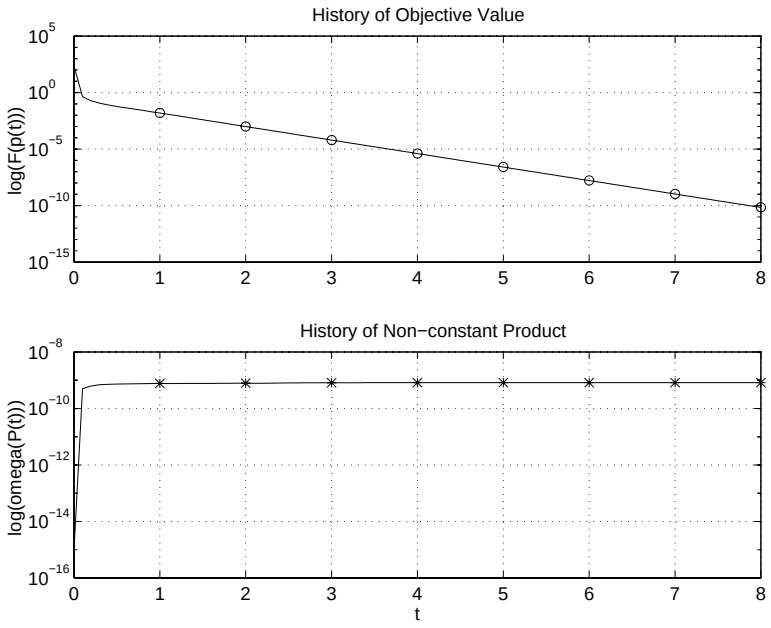

FIG. 4. *History of $F(P(t))$ Example 2 reaching a global solution.*

Using differential system (30) with initial value

$$P(0) = \begin{bmatrix} 2.4660 & -1.0824 & 0.2591 & -0.2234 & -0.4106 \\ 0 & 2.0849 & 0.6624 & 0.2796 & 0.1713 \\ 0 & 0 & 1.3988 & 1.3061 & -0.3510 \\ 0 & 0 & 0 & 1.5571 & 0.1704 \\ 0 & 0 & 0 & 0 & 0.6877 \end{bmatrix}$$

which comes from the Cholesky decomposition of $\tilde{B}$ (see (29)), we calculate the flow $P(t)$. At convergence we convert $P(\infty)$ into $X(\infty)$ and obtain

$$X(\infty) \approx \begin{bmatrix} 7.1728 & 1.8576 & -2.1532 & 0.6710 & 0.5092 \\ 1.8576 & -0.0080 & 0.5149 & 2.1096 & -1.3318 \\ -2.1532 & 0.5149 & -0.9992 & -0.4591 & 0.3602 \\ 0.6710 & 2.1096 & -0.4591 & -3.9520 & -1.2334 \\ 0.5092 & -1.3318 & 0.3603 & -1.2334 & -2.0060 \end{bmatrix}.$$

We note that the off-diagonal elements of $X(\infty)$ agree with those of $\tilde{A}$ up to the integration error. Therefore the local solution $\mathrm{diag}(X(\infty))$ we have found is also a global solution. It is interesting to note that $\mathrm{diag}(X(\infty)) \neq \mathrm{diag}(\hat{A})$, indicating that Problem 3 may have multiple solutions. The history of convergence is in Figure 4.

Theoretically, it should that $P(t)^T P(t) = \tilde{B}$ for all t. Numerical calculation introduces errors. For this reason, we closely watch for the the values of

$$\omega(P(t)) := \|P(t)^T P(t) - \tilde{B}\|. \tag{31}$$

The second graph in Figure 4 indicates that the discrepancy between theoretical expectation and numerical computation is within our tolerance.

Example 3. We want to stress that the optimization problem (5) is non-linear and non-convex. Generally, we cannot expect from any method the luck of hitting the *global* minimizer of any non-linear or non-convex optimization problem by one random starting point. One nice feature of our approach, however, is that we are guaranteed to find a local minimizer regardless where we start and that we have plenty choices of starting points. While it would be nicer to be able to foretell which point/region would serve better as a starting value than the other, the success of such an exploration perhaps is too much to expect for due to the non-linear and non-convex nature of the problem. On the other hand, since we literally can start from anywhere (e.g., any orthogonal matrix in (29)), we find it is possible, though not the best way, to fish for a "better" starting point by trials and errors. We obtain the following results from such a procedure. We have performed many other tests (for the case where a global solution is known to exist) and are always able to find the appropriate starting points after several trials. We have written our code with the convenience of repeated experiments in mind and will make it available upon request.

We report below a case that we think is more challenging than most of the other cases we have tested. Suppose we repeat the experiment in Example 2 with the test data

$$\tilde{A} = \begin{bmatrix} 1.4637 & -0.3440 & 0.6314 & 0.3603 & 1.2990 \\ -0.3440 & -4.1759 & -0.0370 & 0.8424 & -2.5164 \\ 0.6314 & -0.0370 & -0.5261 & 3.1094 & -0.2112 \\ 0.3603 & 0.8424 & 3.1094 & 2.6428 & -0.9722 \\ 1.2990 & -2.5164 & -0.2112 & -0.9722 & -2.0921 \end{bmatrix}$$

$$\tilde{B} = \begin{bmatrix} 2.1437 & 1.7880 & -0.1595 & 0.7567 & -0.0391 \\ 1.7880 & 7.3264 & -2.8274 & -0.0856 & -0.0528 \\ -0.1595 & -2.8274 & 3.8262 & -1.8245 & -1.7653 \\ 0.7567 & -0.0856 & -1.8245 & 5.0857 & 0.4600 \\ -0.0391 & -0.0528 & -1.7653 & 0.4600 & 1.5725 \end{bmatrix}$$

and the target eigenvalues $\sigma(\tilde{A}, \tilde{B}) = \{2.4562, 1.3627, -0.2342, -0.4489, -250.9816\}$. This time the ratio of the eigenvalues of the largest magnitude to the smallest is relatively large and we expect difficulty.

Suppose we start with the upper triangular matrix in the Cholesky decomposition of $\tilde{B}$, i.e, suppose we choose $U = I$ in (29). At convergence we obtain

$$X(\infty) \approx \begin{bmatrix} 2.9383 & -0.3450 & 0.6401 & 0.3600 & 1.2989 \\ -0.3450 & -13.6834 & -0.0605 & 0.8413 & -2.5144 \\ 0.6401 & -0.0605 & -0.2814 & 2.9862 & -0.2303 \\ 0.3600 & 0.8413 & 2.9862 & 0.3243 & -0.9734 \\ 1.2989 & -2.5144 & -0.2303 & -0.9734 & 0.1012 \end{bmatrix}.$$

Note that the off-diagonal elements of $X(\infty)$ are close, but not within the expected integration error, to those of $\tilde{A}$. From Figure 5 we are convinced that we have reached

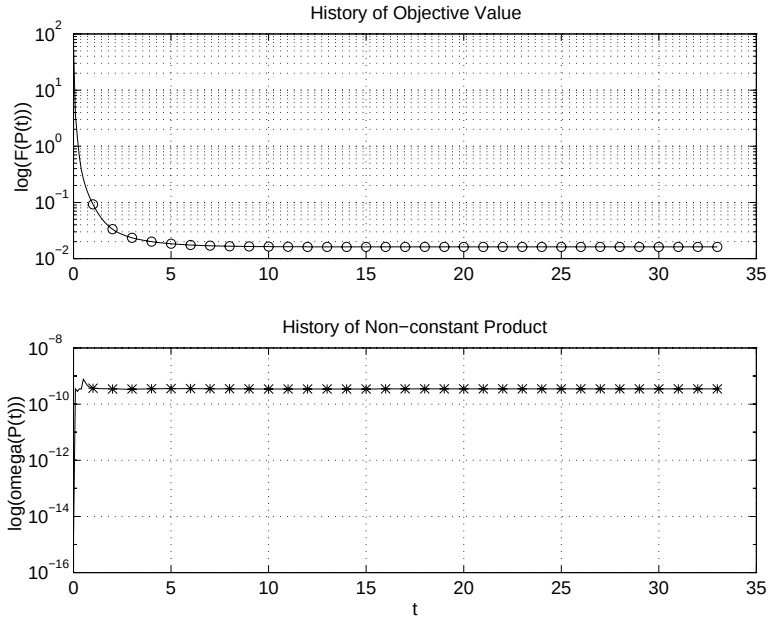

FIG. 5. *History of $F(P(t))$ in Example 3 reaching a local solution.*

only a local solution, although that solution is quite close to a global solution. We have checked that $\sigma(X(\infty), \tilde{B})$ agrees with $\sigma(\tilde{A}, \tilde{B})$ up to the integration error.

This example illustrate another difficulty associated with Problem 3. We know that in Problem 3 only the diagonal elements of $\tilde{A}$ are allowed to vary. The off-diagonal elements of $\tilde{A}$ are not supposed to change, but we find that is not the case in our $X(\infty)$. Suppose we project $X(\infty)$ down to the affine subspace of $\tilde{A}$ plus all diagonal matrices to maintain the off-diagonal elements. The eigenvalues of the corresponding projected pair are given by

$$\sigma(\text{off-diag}(\tilde{A}) + \text{diag}(X(\infty)), \tilde{B}) = \{2.4535, 1.4392, -0.2210, -0.4673, -245.6114\}.$$

These values again are close but not within the integration error to the desired target eigenvalues. In other words, this example demonstrates a case where the spectral constraint and the structural constraint cannot be satisfied simultaneously by a local solution.

Suppose we change the starting value to

$$P(0) = \begin{bmatrix} -0.4186 & 0.4414 & -0.7581 & 1.3847 & -0.1868 \\ 0.3510 & 0.1044 & 0.8450 & -0.8140 & -0.5692 \\ -0.6032 & -2.4090 & 0.3297 & 0.3488 & 0.4427 \\ -1.1340 & -0.7609 & 0.9793 & -1.2783 & -0.6188 \\ 0.4421 & -0.8593 & 1.2123 & 0.8660 & -0.7967 \end{bmatrix}$$

which is obtained by multiplying a specific orthogonal matrix (acquired by random trials) to the upper triangular matrix in the Cholesky decomposition of $\tilde{B}$ (see (29)) .

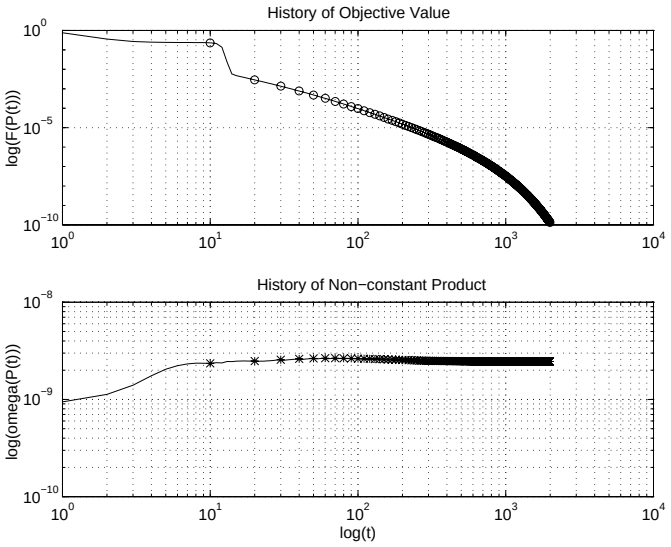

FIG. 6. *History of $F(P(t))$ in Example 3 reaching a global solution.*

It turns out that we are able to find a global solution

$$X(\infty) = \begin{bmatrix} 1.4238 & -0.3440 & 0.6315 & 0.3603 & 1.2990 \\ -0.3440 & -4.1785 & -0.0370 & 0.8424 & -2.5164 \\ 0.6315 & -0.0370 & -0.8734 & 3.1093 & -0.2113 \\ 0.3603 & 0.8424 & 3.1093 & 3.0295 & -0.9722 \\ 1.2990 & -2.5164 & -0.2113 & -0.9722 & -1.8037 \end{bmatrix}$$

that satisfies both the spectral and the structural constraints. The history of integration is plotted in Figure 6. The much longer length of integration required for convergence perhaps is due to the stiffness.

6. Conclusion. We have proposed a general framework for the least squares approximation of symmetric-definite pencils subject to generalized eigenvalue constraints. We have illustrated how this approach can be adapted to different applications, including the inverse generalized eigenvalue problems. Although Problem 2 has already enjoyed efficient and reliable numerical algorithms. There are few methods available for Problem 1 and Problem 3. Our approach unifies these different problems under the same framework. The versatility of our method by specifying V_1 and V_2 seem quite interesting.

We have experimented with several descent flows proposed in this paper by using available ordinary differential equation solvers. Our methods guarantee the global convergence to a local solution. By changing integral paths, a global solution sometimes can be reached. It remains to be studied whether a special-purpose integrator/implementation can be developed to make our approach more efficient.

REFERENCES

[1] S. Boyd, L. E. Ghaoui, E. Feron and V. Balakrishnan, *Linear Matrix inequalities in System and Control Theory*, 15, Studies in Applied Mathematics, SIAM, Philadelphia, 1994.

[2] X. Chen and M. T. Chu, *On the least squares solution of inverse eigenvalue problems*, SIAM J. Numer. Anal., to appear, 1996.

[3] M. T. Chu, *A continuous approximation to the generalized Schur decomposition*, Linear Alg. Appl., 78(1986), 119-132.

[4] M. T. Chu and K. R. Driessel, *The projected gradient method for least squares matrix approximations with spectral constraints*, SIAM J. Numer. Anal., 27(1990), pp. 1050-1060.

[5] M. T. Chu, *A continuous Jacobi-like approach to the simultaneous reduction of real matrices*, Linear Alg. Appl., 147(1991), pp. 75-96.

[6] J. W. Demmel and A. Edelman, *The dimension of matrices (matrix pencils) with given Jordan (Kronecker) canonical forms*, Linear Alg. Appl., 230(1995), pp. 47-60.

[7] G. H. Golub and C. F. Van Loan, *Matrix Computations*, second ed., The Johns Hopkins University Press, Baltimore, MA, 1989.

[8] S. Friedland, J. Nocedal and M. L. Overton, *The formulation and analysis of numerical methods for inverse eigenvalue problems*, SIAM J. Numer. Anal. 24(1987), pp. 634-667.

[9] C. C. Gonzaga, *Path-following methods for linear programming*, SIAM Rev., 34(1992), pp. 167-224.

[10] A. J. Hoffman and H. Wielandt, *The variation of the spectrum of a normal matrix*, Duke Math. J., 20(1953), pp. 37-39.

[11] N. G. Lloyd, *Degree Theory*, Cambridge University Press, New York, 1978.

[12] J. Milnor, *Singular Points of Complex Hyper surfaces*, Annals of Mathematics Studies 61, Princeton University Press, Princeton, New Jersey, 1968.

[13] Yu. Nesterov and A. Nemirovsky, *Interior-point polynomial methods in convex programming*, Vol. 13, in Studies in Applied Math., SIAM, Philadelphia, 1994.

[14] J. D. Pryce, *Numerical Solution of Sturm-Liouville Problems*, Oxford University Press, New York, 1993.

[15] L. F. Shampine and M. K. Gordon, *Computer Solution of Ordinary Differential Equations: The Initial Value Problem*, Freeman, San Francisco, 1975.

[16] L. F. Shampine and M. W. Reichelt, *The MATLAB ODE suite*, preprint, 1995, (available from *ftp.mathworks.com* in the directory *pub/mathworks/toolbox/matlab/funfun*.)

[17] G. W. Stewart and J. G. Sun, *Matrix Perturbation Theory*, Academic Press, San Diego, CA, 1990.

[18] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38(1996), 49-95.

[19] M. H. Wright, *Interior methods for constrained optimization*, in Acta Numerica 1992, A. Iserles, ed., Cambridge University Press, New York, 1992, pp. 341-407.

[20] M. H. Wright, *Some properties of the Hessian of the logarithmic barrier function*, Math. Programming, 67(1994), pp. 265-295.