© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Chapter Two

Motivation and Applications

The problem of optimizing a real-valued function on a matrix manifold appears in a wide variety of computational problems in science and engineering. In this chapter we discuss several examples that provide motivation for the material presented in later chapters. In the first part of the chapter, we focus on the eigenvalue problem. This application receives special treatment because it serves as a running example throughout the book. It is a problem of unquestionable importance that has been, and still is, extensively researched. It falls naturally into the geometric framework proposed in this book as an optimization problem whose natural domain is a matrix manifold—the underlying symmetry is related to the fact that the notion of an eigenvector is scale-invariant. Moreover, there are a wide range of related problems (eigenvalue decompositions, principal component analysis, generalized eigenvalue problems, etc.) that provide a rich collection of illustrative examples that we will use to demonstrate and compare the techniques proposed in later chapters.

Later in this chapter, we describe several research problems exhibiting promising symmetry to which the techniques proposed in this book have not yet been applied in a systematic way. The list is far from exhaustive and is very much the subject of ongoing research. It is meant as an invitation to the reader to consider the broad scope of computational problems that can be cast as optimization problems on manifolds.

2.1 A CASE STUDY: THE EIGENVALUE PROBLEM

The problem of computing eigenspaces and eigenvalues of matrices is ubiquitous in engineering and physical sciences. The general principle of computing an eigenspace is to reduce the complexity of a problem by focusing on a few relevant quantities and dismissing the others. Eigenspace computation is involved in areas as diverse as structural dynamics [GR97], control theory [PLV94], signal processing [CG90], and data mining [BDJ99]. Considering the importance of the eigenproblem in so many engineering applications, it is not surprising that it has been, and still is, a very active field of research.

Let $\mathbb{F}$ stand for the field of real or complex numbers. Let A be an $n \times n$ matrix with entries in $\mathbb{F}$. Any nonvanishing vector $v \in \mathbb{C}^n$ that satisfies

$$Av = \lambda v$$

for some $\lambda \in \mathbb{C}$ is called an *eigenvector* of A; λ is the associated *eigen-*

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

value, and the couple (λ, v) is called an *eigenpair*. The set of eigenvalues of A is called the *spectrum* of A. The eigenvalues of A are the zeros of the *characteristic polynomial* of A,

$$\mathcal{P}_A(z) \equiv \det(A - zI),$$

and their *algebraic multiplicity* is their multiplicity as zeros of $\mathcal{P}_A$. If T is an invertible matrix and (λ, v) is an eigenpair of A, then (λ, Tv) is an eigenpair of TAT^{-1}. The transformation $A \mapsto TAT^{-1}$ is called a *similarity transformation* of A.

A *(linear) subspace* $\mathcal{S}$ of $\mathbb{F}^n$ is a subset of $\mathbb{F}^n$ that is closed under linear combinations, i.e.,

$$\forall x, y \in \mathcal{S},\ \forall a, b \in \mathbb{F} : (ax + by) \in \mathcal{S}.$$

A set $\{y_1, \ldots, y_p\}$ of elements of $\mathcal{S}$ such that every element of $\mathcal{S}$ can be written as a linear combination of $y_1, \ldots, y_p$ is called a *spanning set* of $\mathcal{S}$; we say that $\mathcal{S}$ is the *column space* or simply the *span* of the $n \times p$ matrix $Y = [y_1, \ldots, y_p]$ and that Y *spans* $\mathcal{S}$. This is written as

$$\mathcal{S} = \text{span}(Y) = \{Yx : x \in \mathbb{F}^p\} = Y\mathbb{F}^p.$$

The matrix Y is said to have full (column) rank when the columns of Y are linearly independent, i.e., $Yx = 0$ implies $x = 0$. If Y spans $\mathcal{S}$ and has full rank, then the columns of Y form a *basis* of $\mathcal{S}$. Any two bases of $\mathcal{S}$ have the same number of elements, called the *dimension* of $\mathcal{S}$. The set of all p-dimensional subspaces of $\mathbb{F}^n$, denoted by $\text{Grass}(p, n)$, plays an important role in this book. We will see in Section 3.4 that $\text{Grass}(p, n)$ admits a structure of manifold called the *Grassmann manifold*.

The kernel $\ker(B)$ of a matrix B is the subspace formed by the vectors x such that $Bx = 0$. A scalar λ is an eigenvalue of a matrix A if and only if the dimension of the kernel of $(A - \lambda I)$ is greater than zero, in which case $\ker(A - \lambda I)$ is called the *eigenspace of A related to λ*.

An $n \times n$ matrix A naturally induces a mapping on $\text{Grass}(p, n)$ defined by

$$\mathcal{S} \in \text{Grass}(p, n) \ \mapsto\ A\mathcal{S} := \{Ay : y \in \mathcal{S}\}.$$

A subspace $\mathcal{S}$ is said to be an *invariant subspace* or *eigenspace* of A if $A\mathcal{S} \subseteq \mathcal{S}$. The *restriction* $A|_{\mathcal{S}}$ of A to an invariant subspace $\mathcal{S}$ is the operator $x \mapsto Ax$ whose domain is $\mathcal{S}$. An invariant subspace $\mathcal{S}$ of A is called *spectral* if, for every eigenvalue λ of $A|_{\mathcal{S}}$, the multiplicities of λ as an eigenvalue of $A|_{\mathcal{S}}$ and as an eigenvalue of A are identical; equivalently, $X^T AX$ and $X_\perp^T AX_\perp$ have no eigenvalue in common when $[X|X_\perp]$ satisfies $[X|X_\perp]^T[X|X_\perp] = I_n$ and $\text{span}(X) = \mathcal{S}$.

In many (arguably the majority of) eigenproblems of interest, the matrix A is real and symmetric ($A = A^T$). The eigenvalues of an $n \times n$ symmetric matrix A are reals $\lambda_1 \leq \cdots \leq \lambda_n$, and the associated eigenvectors $v_1, \ldots, v_n$ are real and can be chosen *orthonormal*, i.e.,

$$v_i^T v_j = \begin{cases} 1 \text{ if } i = j, \\ 0 \text{ if } i \neq j. \end{cases}$$

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Equivalently, for every symmetric matrix A, there is an orthonormal matrix V (whose columns are eigenvectors of A) and a diagonal matrix Λ such that $A = V\Lambda V^T$. The eigenvalue λ_1 is called the *leftmost eigenvalue* of A, and an eigenpair (λ_1, v_1) is called a *leftmost eigenpair*. A p-dimensional *leftmost invariant subspace* is an invariant subspace associated with $\lambda_1, \ldots, \lambda_p$. Similarly, a p-dimensional *rightmost invariant subspace* is an invariant subspace associated with $\lambda_{n-p+1}, \ldots, \lambda_n$. Finally, *extreme eigenspaces* refer collectively to leftmost and rightmost eigenspaces.

Given two $n \times n$ matrices A and B, we say that (λ, v) is an eigenpair of the *pencil* (A, B) if

$$Av = \lambda Bv.$$

Finding eigenpairs of a matrix pencil is known as the *generalized eigenvalue problem*. The generalized eigenvalue problem is said to be *symmetric / positive-definite* when A is symmetric and B is symmetric positive-definite (i.e., $x^T Bx > 0$ for all nonvanishing x). In this case, the eigenvalues of the pencil are all real and the eigenvectors can be chosen to form a B-orthonormal basis. A subspace $\mathcal{Y}$ is called a *(generalized) invariant subspace* (or a *deflating subspace*) of the symmetric / positive-definite pencil (A, B) if $B^{-1}Ay \in \mathcal{Y}$ for all $y \in \mathcal{Y}$, which can also be written $B^{-1}A\mathcal{Y} \subseteq \mathcal{Y}$ or $A\mathcal{Y} \subseteq B\mathcal{Y}$. The simplest example is when $\mathcal{Y}$ is spanned by a single eigenvector of (A, B), i.e., a nonvanishing vector y such that $Ay = \lambda By$ for some eigenvalue λ. More generally, every eigenspace of a symmetric / positive-definite pencil is spanned by eigenvectors of (A, B). Obviously, the generalized eigenvalue problem reduces to the standard eigenvalue problem when $B = I$.

2.1.1 The eigenvalue problem as an optimization problem

The following result is instrumental in formulating extreme eigenspace computation as an optimization problem. (Recall that $\mathrm{tr}(A)$, the *trace* of A, denotes the sum of the diagonal elements of A.)

Proposition 2.1.1 *Let A and B be symmetric $n \times n$ matrices and let B be positive-definite. Let $\lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of the pencil (A, B). Consider the* generalized Rayleigh quotient

$$f(Y) = \mathrm{tr}(Y^T AY(Y^T BY)^{-1}) \tag{2.1}$$

defined on the set of all $n \times p$ full-rank matrices. Then the following statements are equivalent:

(i) $\mathrm{span}(Y_*)$ *is a leftmost invariant subspace of* (A, B)*;*
(ii) Y_* *is a global minimizer of* (2.1) *over all $n \times p$ full-rank matrices;*
(iii) $f(Y_*) = \sum_{i=1}^{p} \lambda_i$.

Proof. For simplicity of the development we will assume that $\lambda_p < \lambda_{p+1}$, but the result also holds without this hypothesis. Let V be an $n \times n$ matrix for which $V^T BV = I_n$ and $V^T AV = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, where $\lambda_1 \leq \cdots \leq \lambda_n$.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Such a V always exists. Let $Y \in \mathbb{R}^{n\times p}$ and put $Y = VM$. Since $Y^TBY = I_p$, it follows that $M^TM = I_p$. Then

$$\begin{aligned}\operatorname{tr}(Y^TAY) &= \operatorname{tr}(M^T\operatorname{diag}(\lambda_1,\ldots,\lambda_n)M)\\ &= \sum_{i=1}^{n}\lambda_i\sum_{j=1}^{p}m_{ij}^2\\ &= \sum_{j=1}^{p}\left(\lambda_p + \sum_{i=1}^{p}(\lambda_i-\lambda_p)m_{ij}^2 + \sum_{i=p+1}^{n}(\lambda_i-\lambda_p)m_{ij}^2\right)\\ &= \sum_{i=1}^{p}\lambda_i + \sum_{i=1}^{p}(\lambda_p-\lambda_i)\left(1-\sum_{j=1}^{p}m_{ij}^2\right) + \sum_{j=1}^{p}\sum_{i=p+1}^{n}(\lambda_i-\lambda_p)m_{ij}^2.\end{aligned}$$

Since the second and last terms are nonnegative, it follows that $\operatorname{tr}(Y^TAY) \geq \sum_{i=1}^{p}\lambda_i$. Equality holds if and only if the second and last terms vanish. This happens if and only if the $(n-p)\times p$ lower part of M vanishes (and hence the $p\times p$ upper part of M is orthogonal), which means that $Y = VM$ spans a p-dimensional leftmost invariant subspace of (A,B). □

For the case $p = 1$ and $B = I$, and assuming that the leftmost eigenvalue λ_1 of A has multiplicity 1, Proposition 2.1.1 implies that the global minimizers of the cost function

$$f : \mathbb{R}_*^n \to \mathbb{R} : y \mapsto f(y) = \frac{y^TAy}{y^Ty} \tag{2.2}$$

are the points v_1r, $r \in \mathbb{R}_*$, where $\mathbb{R}_*^n$ is $\mathbb{R}^n$ with the origin removed and v_1 is an eigenvector associated with λ_1. The cost function (2.2) is called the *Rayleigh quotient* of A. Minimizing the Rayleigh quotient can be viewed as an optimization problem on a manifold since, as we will see in Section 3.1.1, $\mathbb{R}_*^n$ admits a natural manifold structure. However, the manifold aspect is of little interest here, as the manifold is simply the classical linear space $\mathbb{R}^n$ with the origin excluded.

A less reassuring aspect of this minimization problem is that the minimizers are not isolated but come up as the continuum $v_1\mathbb{R}_*$. Consequently, some important convergence results for optimization methods do not apply, and several important algorithms may fail, as illustrated by the following proposition.

Proposition 2.1.2 *Newton's method applied to the Rayleigh quotient* (2.2) *yields the iteration* $y \mapsto 2y$ *for every* y *such that* $f(y)$ *is not an eigenvalue of* A.

Proof. Routine manipulations yield $\operatorname{grad} f(y) = \frac{2}{y^Ty}(Ay - f(y)y)$ and $\operatorname{Hess} f(y)[z] = \mathrm{D}(\operatorname{grad} f)(y)[z] = \frac{2}{y^Ty}(Az - f(y)z) - \frac{4}{(y^Ty)^2}(y^TAzy + y^TzAy - 2f(y)y^Tzy) = H_yz$, where $H_y = \frac{2}{y^Ty}(A - f(y)I - \frac{2}{y^Ty}(yy^TA + Ayy^T - 2f(y)yy^T)) = \frac{2}{y^Ty}(I - 2\frac{yy^T}{y^Ty})(A - f(y)I)(I - 2\frac{yy^T}{y^Ty})$. It follows that H_y is

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

singular if and only if $f(y)$ is an eigenvalue of A. When $f(y)$ is not an eigenvalue of A, the Newton equation $H_y\eta = -\text{grad}\, f(y)$ admits one and only one solution, and it is easy to check that this solution is $\eta = y$. In conclusion, the Newton iteration maps y to $y + \eta = 2y$. □

This result is not particular to the Rayleigh quotient. It holds for any function f homogeneous of degree zero, i.e., $f(y\alpha) = f(y)$ for all real $\alpha \neq 0$.

A remedy is to restrain the domain of f to some subset $\mathcal{M}$ of $\mathbb{R}^n_*$ so that any ray $y\mathbb{R}_*$ contains at least one and at most finitely many points of $\mathcal{M}$. Notably, this guarantees that the minimizers are isolated. An elegant choice for $\mathcal{M}$ is the unit sphere

$$S^{n-1} := \{y \in \mathbb{R}^n : y^T y = 1\}.$$

Restricting the Rayleigh quotient (2.2) to S^{n-1} gives us a well-behaved cost function with isolated minimizers. What we lose, however, is the linear structure of the domain of the cost function. The goal of this book is to provide a toolbox of techniques to allow practical implementation of numerical optimization methods on nonlinear embedded (matrix) manifolds in order to address problems of exactly this nature.

Instead of restraining the domain of f to some subset of $\mathbb{R}^n$, another approach, which seems *a priori* more challenging but fits better with the geometry of the problem, is to work on a domain where all points on a ray $y\mathbb{R}_*$ are considered just one point. This viewpoint is especially well suited to eigenvector computation since the useful information of an eigenvector is fully contained in its direction. This leads us to consider the set

$$\mathcal{M} := \{y\mathbb{R}_* : y \in \mathbb{R}^n_*\}.$$

Since the Rayleigh quotient (2.2) satisfies $f(y\alpha) = f(y)$, it induces a well-defined function $\tilde{f}(y\mathbb{R}_*) := f(y)$ whose domain is $\mathcal{M}$. Notice that whereas the Rayleigh quotient restricted to S^{n-1} has two minimizers $\pm v_1$, the Rayleigh quotient $\tilde{f}$ has only one minimizer $v_1\mathbb{R}_*$ on $\mathcal{M}$. It is shown in Chapter 3 that the set $\mathcal{M}$, called the *real projective space*, admits a natural structure of *quotient manifold*. The material in later chapters provides techniques tailored to (matrix) quotient manifold structures that lead to practical implementation of numerical optimization methods. For the simple case of a single eigenvector, algorithms proposed on the sphere are numerically equivalent to those on the real-projective quotient space. However, when the problem is generalized to the computation of p-dimensional invariant subspaces, the quotient approach, which leads to the Grassmann manifold, is seen to be the better choice.

2.1.2 Some benefits of an optimization framework

We will illustrate throughout the book that optimization-based eigenvalue algorithms have a number of desirable properties.

An important feature of all optimization-based algorithms is that optimization theory provides a solid framework for the convergence analysis.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Many optimization-based eigenvalue algorithms exhibit almost global convergence properties. This means that convergence to a solution of the optimization problem is guaranteed for almost every initial condition. The property follows from general properties of the optimization scheme and does not need to be established as a specific property of a particular algorithm.

The speed of convergence of the algorithm is also an intrinsic property of optimization-based algorithms. Gradient-based algorithms converge *linearly*; i.e., the contraction rate of the error between successive iterates is asymptotically bounded by a constant $c < 1$. In contrast, Newton-like algorithms have *superlinear* convergence; i.e., the contraction rate asymptotically converges to zero. (We refer the reader to Section 4.3 for details.)

Characterizing the global behavior and the (local) convergence rate of a given algorithm is an important performance measure of the algorithm. In most situations, this analysis is a free by-product of the optimization framework.

Another challenge of eigenvalue algorithms is to deal efficiently with *large-scale* problems. Current applications in data mining or structural analysis easily involve matrices of dimension $10^5 - 10^6$ [AHLT05]. In those applications, the matrix is typically sparse; i.e., the number of nonzero elements is $O(n)$ or even less, where n is the dimension of the matrix. The goal in such applications is to compute a few eigenvectors corresponding to a small relevant portion of the spectrum. Algorithms are needed that require a small storage space and produce their iterates in $O(n)$ operations. Such algorithms permit matrix-vector products $x \mapsto Ax$, which require $O(n)$ operations if A is sparse, but they forbid matrix factorizations, such as QR and LU, that destroy the sparse structure of A. Algorithms that make use of A only in the form of the operator $x \mapsto Ax$ are called *matrix-free*.

All the algorithms in this book, designed and analyzed using a differential geometric optimization approach, satisfy at least some of these requirements. The trust-region approach presented in Chapter 7 satisfies all the requirements. Such strong convergence analysis is rarely encountered in available eigenvalue methods.

2.2 RESEARCH PROBLEMS

This section is devoted to briefly presenting several general computational problems that can be tackled by a manifold-based optimization approach. Research on the problems presented is mostly at a preliminary stage and the discussion provided here is necessarily at the level of an overview. The interested reader is encouraged to consult the references provided.

2.2.1 Singular value problem

The singular value decomposition is one of the most useful tasks in numerical computations [HJ85, GVL96], in particular when it is used in dimension

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

reduction problems such as principal component analysis [JW92].

Matrices U, Σ, and V form a *singular value decomposition* (SVD) of an arbitrary matrix $A \in \mathbb{R}^{m\times n}$ (to simplify the discussion, we assume that $m \geq n$) if

$$A = U\Sigma V^T, \tag{2.3}$$

with $U \in \mathbb{R}^{m\times m}$, $U^TU = I_m$, $V \in \mathbb{R}^{n\times n}$, $V^TV = I_n$, $\Sigma \in \mathbb{R}^{m\times n}$, Σ diagonal with diagonal entries $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$. Every matrix A admits an SVD. The diagonal entries σ_i of Σ are called the singular values of A, and the corresponding columns u_i and v_i of U and V are called the left and right singular vectors of A. The triplets (σ_i, u_i, v_i) are then called singular triplets of A. Note that an SVD expresses the matrix A as a sum of rank-1 matrices,

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T.$$

The SVD is involved in several least-squares problems. An important example is the best low-rank approximation of an $m \times n$ matrix A in the least-squares sense, i.e.,

$$\arg\min_{X\in\mathcal{R}_p} \|A - X\|_F^2,$$

where $\mathcal{R}_p$ denotes the set of all $m \times n$ matrices with rank p and $\|\cdot\|_F^2$ denotes the Frobenius norm, i.e., the sum of the squares of the elements of its argument. The solution of this problem is given by a truncated SVD

$$X = \sum_{i=1}^{p} \sigma_i u_i v_i^T,$$

where (σ_i, u_i, v_i) are singular triplets of A (ordered by decreasing value of σ). This result is known as the Eckart-Young-Mirsky theorem; see Eckart and Young [EY36] or, e.g., Golub and Van Loan [GVL96].

The singular value problem is closely related to the eigenvalue problem. It follows from (2.3) that $A^TA = V\Sigma^2V^T$, hence the squares of the singular values of A are the eigenvalues of A^TA and the corresponding right singular vectors are the corresponding eigenvectors of A^TA. Similarly, $AA^T = U\Sigma^2U^T$, hence the left singular vectors of A are the eigenvectors of AA^T. One approach to the singular value decomposition problem is to rely on eigenvalue algorithms applied to the matrices A^TA and AA^T. Alternatively, it is possible to compute simultaneously a few dominant singular triplets (i.e., those corresponding to the largest singular values) by maximizing the cost function

$$f(U, V) = \mathrm{tr}(U^TAVN)$$

subject to $U^TU = I_p$ and $V^TV = I_p$, where $N = \mathrm{diag}(\mu_1, \ldots, \mu_p)$, with $\mu_1 > \cdots > \mu_p > 0$ arbitrary. If (U, V) is a solution of this maximization problem, then the columns u_i of U and v_i of V are the ith dominant left and right singular vectors of A. This is an optimization problem on a manifold; indeed, constraint sets of the form $\{U \in \mathbb{R}^{n\times p} : U^TU = I_p\}$ have the structure of an embedded submanifold of $\mathbb{R}^{n\times p}$ called the *(orthogonal) Stiefel manifold* (Section 3.3), and the constraint set for (U, V) is then a product manifold (Section 3.1.6).

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

2.2.2 Matrix approximations

In the previous section, we saw that the truncated SVD solves a particular kind of matrix approximation problem, the best low-rank approximation in the least-squares sense. There are several other matrix approximation problems that can be written as minimizing a real-valued function on a manifold.

Within the matrix nearness framework

$$\min_{X \in \mathcal{M}} \|A - X\|_F^2,$$

we have, for example, the following symmetric positive-definite least-squares problem.

$$\begin{aligned} \text{Find} \quad & C \in \mathbb{R}^{n \times n} \\ \text{to minimize} \quad & \|C - C_0\|^2 \\ \text{subject to} \quad & \operatorname{rank}(C) = p,\ C = C^T,\ C \succeq 0, \end{aligned} \tag{2.4}$$

where $C \succeq 0$ denotes that C is positive-semidefinite; i.e., $x^T C x \geq 0$ for all $x \in \mathbb{R}^n$. We can rephrase this constrained problem as a problem on the set $\mathbb{R}_*^{n \times p}$ of all $n \times p$ full-rank matrices by setting $C = YY^T$, $Y \in \mathbb{R}_*^{n \times p}$. The new search space is simpler, but the new cost function

$$f : \mathbb{R}_*^{n \times p} \to \mathbb{R} : Y \mapsto \|YY^T - C_0\|^2$$

has the symmetry property $f(YQ) = f(Y)$ for all orthonormal $p \times p$ matrices Q, hence minimizers of f are not isolated and the problems mentioned in Section 2.1 for Rayleigh quotient minimization are likely to appear. This again points to a quotient manifold approach, where a set $\{YQ : Q^T Q = I\}$ is identified as one point of the quotient manifold.

A variation on the previous problem is the best low-rank approximation of a correlation matrix by another correlation matrix [BX05]:

$$\begin{aligned} \text{Find} \quad & C \in \mathbb{R}^{n \times n} \\ \text{to minimize} \quad & \|C - C_0\|^2 \\ \text{subject to} \quad & \operatorname{rank}(C) = p,\ C_{ii} = 1\ (i = 1, \ldots, n),\ C \succeq 0. \end{aligned} \tag{2.5}$$

Again, setting $C = YY^T$, $Y \in \mathbb{R}_*^{n \times p}$, takes care of the rank constraint. Replacing this form in the constraint $C_{ii} = 1$, $i = 1, \ldots, n$, yields $\operatorname{diag}(YY^T) = I$. This constraint set can be shown to admit a manifold structure called an *oblique manifold*:

$$\mathcal{OB} := \{Y \in \mathbb{R}_*^{n \times p} : \operatorname{diag}(YY^T) = I_n\};$$

see, e.g., [Tre99, TL02, AG06]. This manifold-based approach is further developed in [GP07].

A more general class of matrix approximation problems is the *Procrustes problem* [GD04]

$$\min_{X \in \mathcal{M}} \|AX - B\|_F^2, \quad A \in \mathbb{R}^{l \times m}, B \in \mathbb{R}^{l \times n}, \tag{2.6}$$

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

where $\mathcal{M} \subseteq \mathbb{R}^{m\times n}$. Taking $\mathcal{M} = \mathbb{R}^{m\times n}$ yields a standard least-squares problem. The orthogonal case, $\mathcal{M} = O_n = \{X \in \mathbb{R}^{n\times n} : X^TX = I\}$, has a closed-form solution in terms of the polar decomposition of B^TA [GVL96]. The case $\mathcal{M} = \{X \in \mathbb{R}^{m\times n} : X^TX = I\}$, where $\mathcal{M}$ is a Stiefel manifold, is known as the *unbalanced orthogonal Procrustes problem*; see [EP99] and references therein. The case $\mathcal{M} = \{X \in \mathbb{R}^{n\times n} : \operatorname{diag}(X^TX) = I_n\}$, where $\mathcal{M}$ is an oblique manifold, is called the *oblique Procrustes problem* [Tre99, TL02].

2.2.3 Independent component analysis

Independent component analysis (ICA), also known as blind source separation (BSS), is a computational problem that has received much attention in recent years, particularly for its biomedical applications [JH05]. A typical application of ICA is the "cocktail party problem", where the task is to recover one or more signals, supposed to be statistically independent, from recordings where they appear as linear mixtures. Specifically, assume that n measured signals $x(t) = [x_1(t), \ldots, x_n(t)]^T$ are instantaneous linear mixtures of p underlying, statistically independent source signals $s(t) = [s_1(t), \ldots, s_p(t)]^T$. In matrix notation, we have

$$x(t) = As(t),$$

where the $n \times p$ matrix A is an unknown constant *mixing matrix* containing the mixture coefficients. The ICA problem is to identify the mixing matrix A or to recover the source signals $s(t)$ using only the observed signals $x(t)$.

This problem is usually translated into finding an $n \times p$ *separating matrix* (or *demixing matrix*) W such that the signals $y(t)$ given by

$$y(t) = W^Tx(t)$$

are "as independent as possible". This approach entails defining a cost function $f(W)$ to measure the independence of the signals $y(t)$, which brings us to the realm of numerical optimization. This separation problem, however, has the structural symmetry property that the measure of independence of the components of $y(t)$ should not vary when different scaling factors are applied to the components of $y(t)$. In other words, the cost function f should satisfy the invariance property $f(WD) = f(W)$ for all nonsingular diagonal matrices D. A possible choice for the cost function f is the log likelihood criterion

$$f(W) := \sum_{k=1}^{K} n_k(\log\det\operatorname{diag}(W^*C_kW) - \log\det(W^*C_kW)), \tag{2.7}$$

where the C_k's are covariance-like matrices constructed from $x(t)$ and $\operatorname{diag}(A)$ denotes the diagonal matrix whose diagonal is the diagonal of A; see, e.g., [Yer02] for the choice of the matrices C_k, and [Pha01] for more information on the cost function (2.7).

The invariance property $f(WD) = f(W)$, similarly to the homogeneity property observed for the Rayleigh quotient (2.2), produces a continuum of

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

minimizers if W is allowed to vary on the whole space of $n \times p$ matrices. Much as in the case of the Rayleigh quotient, this can be addressed by restraining the domain of f to a constraint set that singles out finitely many points in each equivalence class $\{WD : D \text{ diagonal}\}$; a possible choice for the constraint set is the oblique manifold

$$\mathcal{OB} = \{W \in \mathbb{R}_*^{n \times p} : \operatorname{diag}(WW^T) = I_n\}.$$

Another possibility is to identify all the matrices within an equivalence class $\{WD : D \text{ diagonal}\}$ as a single point, which leads to a quotient manifold approach.

Methods for ICA based on differential-geometric optimization have been proposed by, among others, Amari *et al.* [ACC00], Douglas [Dou00], Rahbar and Reilly [RR00], Pham [Pha01], Joho and Mathis [JM02], Joho and Rahbar [JR02], Nikpour *et al.* [NMH02], Afsari and Krishnaprasad [AK04], Nishimori and Akaho [NA05], Plumbley [Plu05], Absil and Gallivan [AG06], Shen *et al.* [SHS06], and Hüeper *et al.* [HSS06]; see also several other references therein.

2.2.4 Pose estimation and motion recovery

In the pose estimation problem, an object is known via a set of landmarks $\{m_i\}_{i=1,\dots,N}$, where $m_i := (x_i, y_i, z_i)^T \in \mathbb{R}^3$ are the three coordinates of the ith landmark in an object-centered frame. The coordinates m_i' of the landmarks in a camera-centered frame obey a rigid body displacement law

$$m_i' = Rm_i + t,$$

where $R \in SO_3$ (i.e., $R^TR = I$ and $\det(R) = 1$) represents a rotation and $t \in \mathbb{R}^3$ stands for a translation. Each landmark point produces a normalized image point in the image plane of the camera with coordinates

$$u_i = \frac{Rm_i + t}{e_3^T(Rm_i + t)}.$$

The pose estimation problem is to estimate the pose (R, t) in the manifold $SO_3 \times \mathbb{R}^3$ from a set of point correspondences $\{(u_i, m_i)\}_{i=1,\dots,N}$. A possible approach is to minimize the real-valued function

$$f : SO_3 \times \mathbb{R}^3 \to \mathbb{R} : (R, t) \mapsto \sum_{i=1}^{N} \|(I - u_iu_i^T)(Rm_i + t)\|^2,$$

which vanishes if and only if the points u_i and m_i' are collinear, i.e., u_i is indeed the coordinate vector of the projection of the ith landmark onto the image plane of the camera. This is an optimization problem on the manifold $SO_3 \times \mathbb{R}^3$. Since rigid body motions can be composed to obtain another rigid body motion, this manifold possesses a group structure called the *special Euclidean group* SE_3.

A related problem is motion and structure recovery from a sequence of images. Now the object is unknown, but two or more images are available from

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

different angles. Assume that N landmarks have been selected on the object and, for simplicity, consider only two images of the object. The coordinates m_i' and m_i'' of the ith landmark in the first and second camera frames are related by a rigid body motion

$$m_i'' = Rm_i' + t.$$

Again without loss of generality, the coordinates of the projections of the ith landmark onto each camera image plane are given by $p_i = \frac{m_i'}{e_3^T m_i'}$ and $q_i = \frac{m_i''}{e_3^T m_i''}$. The motion and structure recovery problem is, from a set of corresponding image points $\{(p_i, q_i)\}_{i=1,\dots,N}$, to recover the camera motion (R, t) and the three-dimensional coordinates of the points that the images correspond to. It is a classical result in computer vision that corresponding coordinate vectors p and q satisfy the *epipolar constraint*

$$p^T R^T t^\wedge q = 0,$$

where $t^\wedge$ is the 3×3 skew-symmetric matrix

$$t^\wedge := \begin{bmatrix} 0 & -t_3 & t_2 \\ t_3 & 0 & -t_1 \\ -t_2 & t_1 & 0 \end{bmatrix}.$$

To recover the motion $(R, t) \in SO_3 \times \mathbb{R}^3$ from a given set of image correspondences $\{(p_i, q_i)\}_{i=1,\dots,N}$, it is thus natural to consider the cost function

$$f(R, t) := \sum_{i=1}^{N} (p_i^T R^T t^\wedge q_i)^2, \quad p_i, q_i \in \mathbb{R}^3, (R, t) \in SO_3 \times \mathbb{R}^3.$$

This function is homogeneous in t. As in the case of Rayleigh quotient minimization, this can be addressed by restricting t to the unit sphere S^2, which yields the problem of minimizing the cost function

$$f(R, t) := \sum_{i=1}^{N} (p_i^T R^T t^\wedge q_i)^2, \quad p_i, q_i \in \mathbb{R}^3, (R, t) \in SO_3 \times S^2.$$

Equivalently, this problem can be written as the minimization of the cost function

$$f(E) =: \sum_{i=1}^{N} (p_i^T E q_i)^2, \quad p_i, q_i \in \mathbb{R}^3, E \in \mathcal{E}_1,$$

where $\mathcal{E}_1$ is the *normalized essential manifold*

$$\mathcal{E}_1 := \{Rt^\wedge : R \in SO_3,\ t^\wedge \in \mathfrak{so}_3,\ \tfrac{1}{2}\mathrm{tr}((t^\wedge)^T t^\wedge) = 1\}.$$

($\mathfrak{so}_3 = \{\Omega \in \mathbb{R}^{3\times 3} : \Omega^T = -\Omega\}$ is the Lie algebra of SO_3, and the tr function returns the sum of the diagonal elements of its argument.)

For more details on multiple-view geometry, we refer the reader to Hartley and Zisserman [HZ03]. Applications of manifold optimization to computer vision problems can be found in the work of Ma *et al.* [MKS01], Lee and Moore [LM04], Liu *et al.* [LSG04], and Helmke *et al.* [HHLM07].

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

2.3 NOTES AND REFERENCES

Each chapter of this book (excepting the introduction) has a Notes and References section that contains pointers to the literature. In the following chapters, all the citations will appear in these dedicated sections.

Recent textbooks and surveys on the eigenvalue problem include Golub and van der Vorst [GvdV00], Stewart [Ste01], and Sorensen [Sor02]. An overview of applications can be found in Saad [Saa92]. A major reference for the symmetric eigenvalue problem is Parlett [Par80]. The characterization of eigenproblems as minimax problems goes back to the time of Poincaré. Early references are Fischer [Fis05] and Courant [Cou20], and the results are often referred to as the Courant-Fischer minimax formulation. The formulation is heavily exploited in perturbation analysis of Hermitian eigenstructure. Good overviews are available in Parlett [Par80, §10 and 11, especially §10.2], Horn and Johnson [HJ91, §4.2], and Wilkinson [Wil65, §2]. See also Bhatia [Bha87] and Golub and Van Loan [GVL96, §8.1].

Until recently, the differential-geometric approach to the eigenproblem had been scarcely exploited because of tough competition from some highly efficient mainstream algorithms combined with a lack of optimization algorithms on manifolds geared towards computational efficiency. However, thanks in particular to the seminal work of Helmke and Moore [HM94] and Edelman, Arias, and Smith [Smi93, Smi94, EAS98], and more recent work by Absil *et al.* [ABG04, ABG07], manifold-based algorithms have now appeared that are competitive with state-of-the-art methods and sometimes shed new light on their properties. Papers that apply differential-geometric concepts to the eigenvalue problem include those by Chen and Amari [CA01], Lundström and Eldén [LE02], Simoncinin and Eldén [SE02], Brandts [Bra03], Absil *et al.* [AMSV02, AMS04, ASVM04, ABGS05, ABG06b], and Baker *et al.* [BAG06]. One "mainstream" approach capable of satisfying all the requirements in Section 2.1.2 is the Jacobi-Davidson conjugate gradient (JDCG) method of Notay [Not02]. Interestingly, it is closely related to an algorithm derived from a manifold-based trust-region approach (see Chapter 7 or [ABG06b]).

The proof of Proposition 2.1.1 is adapted from [Fan49]. The fact that the classical Newton method fails for the Rayleigh quotient (Proposition 2.1.2) was pointed out in [ABG06b], and a proof was given in [Zho06].

Major references for Section 2.2 include Helmke and Moore [HM94], Edelman *et al.* [EAS98], and Lippert and Edelman [LE00]. The cost function suggested for the SVD (Section 2.2.1) comes from Helmke and Moore [HM94, Ch. 3]. Problems (2.4) and (2.5) are particular instances of the *least-squares covariance adjustment problem* recently defined by Boyd and Xiao [BX05]; see also Manton *et al.* [MMH03], Grubisic and Pietersz [GP07], and several references therein.

For general queries, contact webmaster@press.princeton.edu