Eigenvalues and diagonalization

Patrick Breheny

November 15

Introduction

- The next topic in our course, *principal components analysis*, revolves around decomposing large, complicated, multivariate relationships into simple uncorrelated elements
- Today's lecture is a primer on the mathematical tools that provide the basis for these methods
- The proofs of most of today's results are beyond the scope of this course, but the results themselves we will use going forward to derive statistical methods and establish theoretical results

Direction and length

- One can think about decomposing a vector $\mathbf{v}$ into two separate pieces of information, its direction $\mathbf{d}$ and its length λ:

$$\lambda = \|\mathbf{v}\| = \sqrt{\sum_j v_j^2}$$

$$\mathbf{d} = \frac{\mathbf{v}}{\lambda}$$

- This makes it easier to work with the magnitude of a vector (λ is a scalar) while ignoring its direction, and vice versa (*e.g.*, rotating a vector)

Eigenvalues and eigenvectors

- This idea can be extended to matrices as well – after all, what is a matrix $\mathbf{A}$ but a collection of vectors?
- This is the idea behind eigenvalues and eigenvectors, which are defined according to this equation:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

- Any vector of unit length $\mathbf{v}$ which satisfies this equation is special to $\mathbf{A}$: when $\mathbf{A}$ operates in that direction, it acts merely to elongate or shrink
- In other words, the vector is not rotated, and direction is preserved

Eigenvalues and eigenvectors (cont'd)

- The $\mathbf{v}$'s which satisfy this equation are called *eigenvectors*
- The λ's are the *eigenvalues*
- We will begin by considering the case where $\mathbf{A}$ is symmetric; later in the lecture, we will consider more general cases

Number of solutions

- **Theorem:** If $\mathbf{A}$ is a symmetric $n \times n$ matrix, then it has exactly n pairs $\{\lambda_j, \mathbf{v}_j\}$ which satisfy the equation

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

- The n eigenvalues are sometimes referred to as the *spectrum* of $\mathbf{A}$

Eigenvalues, traces, and determinants

- **Theorem:** If $\mathbf{A}$ has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$,

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$$

- **Theorem:** If $\mathbf{A}$ has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$,

$$|\mathbf{A}| = \prod_{i=1}^{n} \lambda_i$$

Eigenvalues and positive definite matrices

- **Theorem:**

  $\mathbf{A}$ positive definite $\Leftrightarrow$ all eigenvalues of $\mathbf{A}$ are positive

- **Theorem:**

  $\mathbf{A}$ positive semidefinite $\Leftrightarrow$ all eigenvalues of $\mathbf{A}$ are nonnegative

Decomposition

- Perhaps the most useful aspect of eigenvalues is that they can be used to factor a matrix into simpler elements
- This general notion is referred to as *matrix factorization* or *matrix decomposition*, and it is the main idea behind principal component analysis
- We introduce now a factorization for symmetric matrices called the *eigendecomposition*; a more general factorization called the singular value decomposition will be introduced later and applies to all matrices

Eigendecomposition

- **Lemma:** If $\mathbf{A}$ is a symmetric matrix, its eigenvectors are orthonormal.
- **Theorem:** Any symmetric matrix $\mathbf{A}$ can be factored into:

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T,$$

where $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues of $\mathbf{A}$, and the columns of $\mathbf{Q}$ contain its orthonormal eigenvectors

Inverses

- **Theorem:** Suppose $\mathbf{A}$ has eigenvectors $\mathbf{Q}$ and eigenvalues $\{\lambda_i\}$. Then $\mathbf{A}^{-1}$ has eigenvectors $\mathbf{Q}$ and eigenvalues $\{\lambda_i^{-1}\}$
- In other words, if $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$,

$$\mathbf{A}^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^T$$

Rank

- As you may have supposed from the fact that eigenvalues are intricately tied up with determinants and inverses, they are also related to the rank of a matrix
- **Theorem:** Suppose $\mathbf{A}$ has rank r. Then $\mathbf{A}$ has r nonzero eigenvalues, and the remaining $n - r$ eigenvalues are equal to zero.

Eigenvalues in action: Ridge regression

- To get a sense for how these facts are useful in statistics, let's look back at some unproven theorems from our lecture on ridge regression
- **Theorem:** For any design matrix $\mathbf{X}$, the quantity $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ is always invertible; thus, there is always a unique solution $\widehat{\boldsymbol{\beta}}^{\text{ridge}}$
- **Theorem:** The degrees of freedom for a ridge regression estimate are

$$\sum \frac{\lambda_i}{\lambda_i + \lambda},$$

where $\{\lambda_i\}$ are the eigenvalues of $\mathbf{X}^T\mathbf{X}$

Diagonalization

- We have seen that eigenvalues are useful because they allow you, when performing certain tasks, to do away with all the complexities of matrix algebra and work on the more familiar level of scalars
- This was possible because all the relevant computations took place on a diagonal matrix, $\mathbf{\Lambda}$
- This idea is known as *diagonalization*
- Perhaps the most amazing result in all of linear algebra is that any matrix can be diagonalized in what is known as the *singular value decomposition*

The singular value decomposition

- **Theorem:** For any matrix $\mathbf{A}$, we can write

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T,$$

  where $\mathbf{D}$ is a diagonal matrix with nonnegative entries, and both $\mathbf{U}$ and $\mathbf{V}$ are orthogonal (*i.e.*, $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$).
- Non-square matrices do not have eigenvalues, but the elements of $\mathbf{D}$ (called the *singular values* of $\mathbf{A}$) are the square roots of the eigenvalues of $\mathbf{A}^T\mathbf{A}$ (or $\mathbf{A}\mathbf{A}^T$)

Dimensions of the SVD

- If $\mathbf{A}$ is an $n \times n$ matrix, then $\mathbf{U}$, $\mathbf{D}$, and $\mathbf{V}$ are all $n \times n$ matrices
- If $\mathbf{A}$ is an $n \times p$ matrix with $n > p$, then $\mathbf{U}$ is $n \times p$, while both $\mathbf{D}$ and $\mathbf{V}$ are $p \times p$ matrices
- If $\mathbf{A}$ is an $n \times p$ matrix with $n < p$, then $\mathbf{V}^T$ is $n \times p$, while both $\mathbf{D}$ and $\mathbf{U}$ are $n \times n$ matrices
- In other words, $\mathbf{D}$ is square, with dimension equal to the minimum of n and p

SVD, applied to ridge regression

- As before, let's go back to ridge regression for an example of the singular value decomposition in action
- **Theorem:** Let $\mathbf{X} = \mathbf{UDV}^T$. Then

$$\begin{aligned}
\mathbf{X}^T\mathbf{X} &= \mathbf{VD}^2\mathbf{V}^T \\
\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} &= \mathbf{UU}^T\mathbf{y} \\
\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} &= \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})\mathbf{V}^T \\
\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} &= \mathbf{USU}^T\mathbf{y},
\end{aligned}$$

where $\mathbf{S}$ is a diagonal matrix with elements

$$S_{jj} = \frac{d_j^2}{d_j^2 + \lambda}$$

Remarks

- Once again we see the shrinkage aspect of ridge regression
- Note, however, that small values of d_j are shrunken quite a lot, while large values are shrunk very little
- What do small values of d_j mean?
- Recall the connection between the singular values and eigenvalues; in particular, if $\mathbf{X}$ has rank r, then $\mathbf{D}$ has r nonzero elements

Remarks (cont'd)

- The preceding remark holds in a more qualitative sense as well; if $\mathbf{X}$ is nearly singular (*i.e.*, has problems with multicollinearity, but is still full rank), then it will have one or more d_j very close to zero
- Thus we have another view of how ridge regression stabilizes estimation in the presence of multicollinearity:

$$\frac{d_j}{d_j^2 + \lambda}$$

is highly variable and approaches being undefined as $d_j \to 0$ if λ is not present, but with ridge regression, the quantity is much more stable

Dimension reduction

- Instead of shrinkage, a different approach is to eliminate all the small singular values
- In other words, suppose that $\mathbf{X}$ contains 10 explanatory variables, but that two of its singular values are near zero
- We can eliminate those two values and deal with a more stable, rank-8 version of $\mathbf{X}$
- This is the main idea behind principal components analysis, and we will get into specifics in the next lecture; for now, let me mention that this idea, of approximating a matrix with a low-rank version of it that hopefully captures all the important information, goes far beyond statistics and is widely used to reduce the storage and computations involved in working with large matrices