Matrix algebra, vector calculus, and Taylor series

Patrick Breheny

September 9, 2024

Introduction

One final lecture of analysis review, in which we will

- Review matrix algebra
- Use it to go over vector calculus
- Use that to introduce multivariate Taylor series expansions, the most important mathematical tool in this course

Linear algebra

- Note: If this material is unfamiliar to you, consult this review
- As we have seen, it is often useful to *transpose* a matrix (switch its rows and columns around); this is denoted with a superscript $^\top$ or an apostrophe $'$:

$$\mathbf{M} = \begin{bmatrix} 3 & 2 \\ 4 & -1 \\ -1 & 2 \end{bmatrix} \qquad \mathbf{M}^\top = \begin{bmatrix} 3 & 4 & -1 \\ 2 & -1 & 2 \end{bmatrix}$$

Linear and quadratic forms

Matrix products involving linear and quadratic forms come up very often in statistics, and it is important to have an intuitive grasp on what they represent:

$$\mathbf{a}^\top \mathbf{x} = \sum_i a_i x_i; \quad \mathbf{1}^\top \mathbf{x} = \sum_i x_i$$

$$\mathbf{A}^\top \mathbf{x} = (\sum_i a_{i1} x_i \quad \cdots \quad \sum_i a_{ik} x_i)^\top$$

$$\mathbf{a}^\top \mathbf{W} \mathbf{x} = \sum_i \sum_j a_i w_{ij} x_j; \quad \mathbf{a}^\top \mathbf{1} \mathbf{x} = \sum_i \sum_j a_i x_j$$

$$(\mathbf{AWB})_{ij} = \sum_k \sum_m a_{ik} w_{km} b_{mj}$$

Inverses

- **Definition:** The *inverse* of an $n \times n$ matrix $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, is the matrix satisfying $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix.
- Note: We're sort of getting ahead of ourselves by saying that $\mathbf{A}^{-1}$ is "the" matrix satisfying $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$, but it is indeed the case that if a matrix has an inverse, the inverse is unique
- Some useful results:

$$
\begin{aligned}
(\mathbf{A} + \mathbf{B})^\top &= \mathbf{A}^\top + \mathbf{B}^\top \\
(\mathbf{A}\mathbf{B})^\top &= \mathbf{B}^\top \mathbf{A}^\top \\
(\mathbf{A}\mathbf{B})^{-1} &= \mathbf{B}^{-1}\mathbf{A}^{-1} \\
(\mathbf{A}^\top)^{-1} &= (\mathbf{A}^{-1})^\top
\end{aligned}
$$

Singular matrices

- However, not all matrices have inverses; for example

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

- There does not exist a matrix such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_2$
- Such matrices are said to be *singular*
- Remark: Only square matrices have inverses; an $n \times m$ matrix $\mathbf{A}$ might, however, have a *left inverse* (satisfying $\mathbf{BA} = \mathbf{I}_m$) or *right inverse* (satisfying $\mathbf{AB} = \mathbf{I}_n$)

Positive definite

- A related notion is that of a "positive definite" matrix, which (at least for us) applies only to symmetric matrices
- **Definition:** A symmetric $n \times n$ matrix $\mathbf{A}$ is said to be *positive definite* if for all $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0 \qquad \text{if } \mathbf{x} \neq 0$$

- The two notions are related: if $\mathbf{A}$ is positive definite, then (a) $\mathbf{A}$ is not singular and (b) $\mathbf{A}^{-1}$ is also positive definite
- If $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$, then $\mathbf{A}$ is said to be *positive semidefinite*
- In statistics, these classifications are particularly important for variance-covariance matrices, which are always positive semidefinite (and positive definite, if they aren't singular)

Square root of a matrix

- These concepts are important with respect to knowing whether a matrix has a "square root"
- **Definition:** An $n \times n$ matrix $\mathbf{A}$ is said to have a *square root* if there exists a matrix $\mathbf{B}$ such that $\mathbf{BB} = \mathbf{A}$.
- **Theorem:** Let $\mathbf{A}$ be a positive semidefinite matrix. Then there exists a unique matrix $\mathbf{A}^{1/2}$ such that $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$.

Rank

- We also need to be familiar with the concept of matrix rank (there are many ways of defining rank; all are equivalent)
- **Definition:** The *rank* of a matrix is the dimension of its largest nonsingular submatrix.
- For example, the following 3×3 matrix is singular, but contains a nonsingular 2×2 submatrix, so its rank is 2:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & \cancel{3} \\ \cancel{2} & \cancel{4} & \cancel{6} \\ 1 & 0 & \cancel{1} \end{bmatrix}$$

- Note that a nonsingular $n \times n$ matrix has rank n, and is said to be *full rank*

Rank and multiplication

- There are many results and theorems involving rank; we're not going to cover them all, but it is important to know that rank cannot be increased through the process of multiplication
- **Theorem:** For any matrices $\mathbf{A}$ and $\mathbf{B}$ with appropriate dimensions, $\text{rank}(\mathbf{AB}) \leq \text{rank}(\mathbf{A})$ and $\text{rank}(\mathbf{AB}) \leq \text{rank}(\mathbf{B})$.
- In particular, $\text{rank}(\mathbf{A}^\top\mathbf{A}) = \text{rank}(\mathbf{A}\mathbf{A}^\top) = \text{rank}(\mathbf{A})$

Expectation and variance

- In addition, we need some results on expected values of vectors and functions of vectors
- First of all, we need to define expectation and variance as they pertain to random vectors
- **Definition:** Let $\mathbf{x} = (X_1 \; X_2 \; \cdots X_d)^\top$ denote a vector of random variables, then $\mathbb{E}(\mathbf{x}) = (\mathbb{E}X_1 \; \mathbb{E}X_2 \; \cdots \mathbb{E}X_d)^\top$. Meanwhile, $\mathbb{V}\mathbf{x}$ is a $d \times d$ matrix:

$$\mathbb{V}\mathbf{x} = \mathbb{E}\{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\} \text{ with elements}$$
$$(\mathbb{V}\mathbf{x})_{ij} = \mathbb{E}\left\{(X_i - \mu_i)(X_j - \mu_j)\right\},$$

where $\mu_i = \mathbb{E}X_i$. The matrix $\mathbb{V}\mathbf{x}$ is referred to as the *variance-covariance matrix* of $\mathbf{x}$.

Linear and quadratic forms

- Letting $\mathbf{A}$ denote a matrix of constants and $\mathbf{x}$ a random vector with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$,

$$\begin{aligned}
\mathbb{E}(\mathbf{A}^\top \mathbf{x}) &= \mathbf{A}^\top \boldsymbol{\mu} \\
\mathbb{V}(\mathbf{A}^\top \mathbf{x}) &= \mathbf{A}^\top \boldsymbol{\Sigma} \mathbf{A} \\
\mathbb{E}(\mathbf{x}^\top \mathbf{A} \mathbf{x}) &= \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu} + \operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}),
\end{aligned}$$

  where $\operatorname{tr}(\mathbf{A}) = \sum_i A_{ii}$ is the trace of $\mathbf{A}$
- Some useful facts about traces:

$$\begin{aligned}
\operatorname{tr}(\mathbf{AB}) &= \operatorname{tr}(\mathbf{BA}) \\
\operatorname{tr}(\mathbf{A} + \mathbf{B}) &= \operatorname{tr}(\mathbf{A}) + \operatorname{tr}(\mathbf{B}) \\
\operatorname{tr}(c\mathbf{A}) &= c \operatorname{tr}(\mathbf{A}) \\
\operatorname{tr}(\mathbf{A}) &= \operatorname{rank}(\mathbf{A}) \quad \text{if } \mathbf{AA} = \mathbf{A}
\end{aligned}$$

Eigendecompositions

- Finally, we'll also take a moment to introduce some facts about eigenvalues
- The most important thing about eigenvalues is that they allow us to "diagonalize" a matrix: if $\mathbf{A}$ is a symmetric $d \times d$ matrix, then it can be factored into:

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

  where $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$ of $\mathbf{A}$ and the columns of $\mathbf{Q}$ are its eigenvectors
- Furthermore, eigenvectors are orthonormal, so we have $\mathbf{Q}^\top\mathbf{Q} = \mathbf{Q}\mathbf{Q}^\top = \mathbf{I}$

Eigenvalues and "size"

- This is very helpful from a conceptual standpoint, as it allows us to separate the "size" of a matrix ($\mathbf{\Lambda}$) from its "direction(s)" ($\mathbf{Q}$)
- For example, we have already seen that one measure of the size of a matrix is based on $\lambda_{\max}$ (for a symmetric matrix, its spectral norm is its largest eigenvalue)
- In addition, the trace and determinant, two other ways of quantifying the "size" of a matrix, are simple functions of the eigenvalues:
  - $\text{tr}(\mathbf{A}) = \sum_i \lambda_i$
  - $|\mathbf{A}| = \prod_i \lambda_i$

Eigenvalues and inverses

- Once one has obtained the eigendecomposition of $\mathbf{A}$, calculating its inverse is straightforward
- If $\mathbf{A}$ is not singular, then $\mathbf{A}^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^\top$; note that since $\mathbf{\Lambda}$ is diagonal, its inverse is trivial to calculate
- Even if $\mathbf{A}$ is singular, we can obtain something called a "generalized inverse": $\mathbf{A}^- = \mathbf{Q}\mathbf{\Lambda}^-\mathbf{Q}^\top$, where $(\mathbf{\Lambda}^-)_{ii} = \lambda_i^{-1}$ if $\lambda_i \neq 0$ and $(\mathbf{\Lambda}^-)_{ii} = 0$ otherwise
- Many other important properties of matrices can be deduced entirely from their eigenvalues:
  - $\mathbf{A}$ is positive definite if and only if $\lambda_i > 0$ for all i
  - $\mathbf{A}$ is positive semidefinite if and only if $\lambda_i \geq 0$ for all i
  - If $\mathbf{A}$ has rank r, then $\mathbf{A}$ has r nonzero eigenvalues and the remaining $d - r$ eigenvalues are zero

Extreme values

- Lastly, there is a connection between a matrix's eigenvalues and the extreme values of its quadratic form
- Let the eigenvalues $\lambda_1, \ldots, \lambda_d$ of $\mathbf{A}$ be ordered from largest to smallest. Over the set of all vectors $\mathbf{x}$ such that $\|\mathbf{x}\|_2 = 1$,

$$\max \mathbf{x}^\top \mathbf{A}\mathbf{x} = \lambda_1$$

and

$$\min \mathbf{x}^\top \mathbf{A}\mathbf{x} = \lambda_d$$

Real-valued functions: Derivative and gradient

- We're now ready to talk about vector calculus, which is extremely important in statistics
- **Definition:** For a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, its *derivative* is the $1 \times d$ row vector

$$\dot{f}(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1} \cdots \frac{\partial f}{\partial x_d} \right]$$

- In statistics, it is generally more common (but not always the case) to use the gradient (also called "denominator layout" or the "Hessian formulation")

$$\nabla f(\mathbf{x}) = \dot{f}(\mathbf{x})^\top;$$

i.e., $\nabla f(\mathbf{x})$ is a $d \times 1$ column vector

Vector-valued functions

- **Definition:** For a function $\mathbf{f} : \mathbb{R}^d \mapsto \mathbb{R}^k$, its *derivative* is the $k \times d$ matrix with ijth element

$$\dot{\mathbf{f}}(\mathbf{x})_{ij} = \frac{\partial f_i(\mathbf{x})}{\partial x_j}$$

- Correspondingly, the gradient is a $d \times k$ matrix:

$$\nabla \mathbf{f}(\mathbf{x}) = \dot{\mathbf{f}}(\mathbf{x})^\top$$

- In our course, this will usually come up in the context of taking second derivatives; however, by the symmetry of second derivatives, we have

$$\nabla^2 f(\mathbf{x}) = \ddot{f}(\mathbf{x})$$

Vector calculus identities

Inner product:	$\nabla_{\mathbf{x}}(\mathbf{A}^\top\mathbf{x}) = \mathbf{A}$
Quadratic form:	$\nabla_{\mathbf{x}}(\mathbf{x}^\top\mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$
Chain rule:	$\nabla_{\mathbf{x}}\mathbf{f}(\mathbf{y}) = \nabla_{\mathbf{x}}\mathbf{y}\nabla_{\mathbf{y}}\mathbf{f}$
Product rule:	$\nabla(\mathbf{f}^\top\mathbf{g}) = (\nabla\mathbf{f})\mathbf{g} + (\nabla\mathbf{g})\mathbf{f}$
Inverse function theorem:	$\nabla_{\mathbf{x}}\mathbf{y} = (\nabla_{\mathbf{y}}\mathbf{x})^{-1}$

Vector calculus identities (row-vector layout)

Inner product:	$D_{\mathbf{x}}(\mathbf{A}\mathbf{x}) = \mathbf{A}$
Quadratic form:	$D_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}^\top \mathbf{x}) = \mathbf{x}^\top(\mathbf{A} + \mathbf{A}^\top)$
Chain rule:	$D_{\mathbf{x}}\mathbf{f}(\mathbf{y}) = D_{\mathbf{y}}\mathbf{f}\, D_{\mathbf{x}}\mathbf{y}$
Product rule:	$D(\mathbf{f}^\top \mathbf{g}) = \mathbf{g}^\top \dot{\mathbf{f}} + \mathbf{f}^\top \dot{\mathbf{g}}$
Inverse function theorem:	$D_{\mathbf{x}}\mathbf{y} = (D_{\mathbf{y}}\mathbf{x})^{-1}$

I don't expect to use these, but for your future reference, here they are

Practice

Exercise: In linear regression, the ridge regression estimator is obtained by minimizing the function

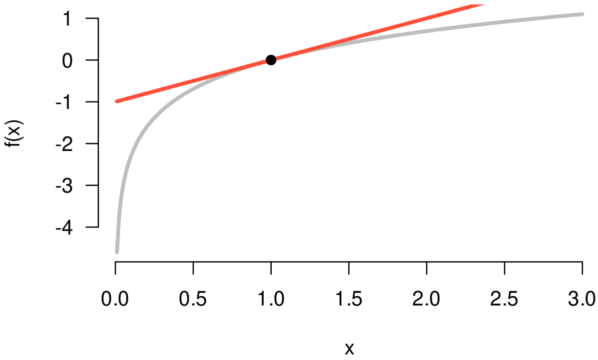
$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2,$$

where λ is a prespecified tuning parameter. Show that

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}.$$

Taylor series: Introduction

As we will see (many times!), it is useful to be able to approximate a complicated function with a simple polynomial (this is the idea behind Taylor series approximation):



Taylor series: Introduction (cont'd)

- It is difficult to overstate the importance of Taylor series expansions to statistical theory, and for that reason we are now going to cover them fairly extensively
- In particular, Taylor's theorem comes in a number of versions, and it is worth knowing several of them, since they come up in statistics quite often
- Furthermore, students often have not seen the multivariate versions of these expansions

Taylor's theorem

- **Theorem (Taylor):** Suppose n is a positive integer and $f : \mathbb{R} \mapsto \mathbb{R}$ is n times differentiable at a point x_0. Then

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + R_n(x, x_0),$$

where the remainder R_n satisfies

$$R_n(x, x_0) = o(|x - x_0|^n) \text{as } x \to x_0$$

- If $f^{(n+1)}(x_0)$ exists, you could also say that R_n is $O(|x - x_0|^{n+1})$
- This form of the remainder is sometimes called the *Peano* form

Taylor's theorem: Lagrange form

- **Theorem (Taylor):** Suppose $f : \mathbb{R} \mapsto \mathbb{R}$ is $n + 1$ times differentiable on an open interval containing x_0. Then for any point x in that interval, there exists $\bar{x} \in (x, x_0)$:

$$R_n(x, x_0) = \frac{f^{(n+1)}(\bar{x})}{(n+1)!}(x - x_0)^{n+1}.$$

- This is also known as the *mean-value form*, as the mean value theorem is the central idea in proving the result

Comparing the two forms

- Comparing the Basic and Lagrange forms for a second-order expansion,

$$f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + o(|x - x_0|^2)$$
$$f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(\bar{x})(x - x_0)^2$$

- We can see that in the second case, we have a simpler expression, but to obtain it, we require f'' to exist along the entire interval from x to x_0, not just at the point x_0

Example: Absolute value

- For example, consider approximating the function $f(x) = |x|$ at $x_0 = -0.1$
- Note that f' exists at x_0, but not at 0
- The basic form of Taylor's theorem says that if we get close enough to x_0, the approximation $f(-0.1) + f'(-0.1)(x + 0.1)$ becomes very accurate – indeed, the remainder is exactly zero for any x within 0.1 of x_0
- However, suppose $x = 0.2$; since f is not differentiable at zero, we are not guaranteed the existence of a point $\bar{x}$ such that

$$f(0.2) = f(-0.1) + 0.3f'(\bar{x});$$

and indeed in this case no such point exists

Lagrange bound

- One reason why the Lagrange form is more powerful is that it allows us to establish error bounds – to know exactly how close x must be to x_0 in order to ensure that the approximation error is less than ϵ
- In particular, if there exists an M such that $\left|f^{(n+1)}(\cdot)\right| \leq M$ over the interval (x, x_0), then

$$|R_n(x)| \leq \frac{M}{(n+1)!} \left|x - x_0\right|^{n+1}$$

Multivariable forms of Taylor's theorem

- We now turn our attention to the multivariate case
- For the sake of clarity, I'll present the first- and second-order expansions for each of the previous forms, rather than abstract formulae involving $f^{(n)}$
- Lastly, I'll provide a form that goes out to third order, although higher orders are less convenient as they can't be represented compactly using vectors and matrices
- Note that these forms are only covering the case of scalar-valued functions $f : \mathbb{R}^d \mapsto \mathbb{R}$; we will need results for the vector-valued case $f : \mathbb{R}^d \mapsto \mathbb{R}^k$ as well, but we will go over that in a later lecture

Taylor's theorem

- **Theorem (Taylor):** Suppose $f : \mathbb{R}^d \mapsto \mathbb{R}$ is differentiable at a point $\mathbf{x}_0$. Then

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0) + o(\|\mathbf{x} - \mathbf{x}_0\|)$$

- **Theorem (Taylor):** Suppose $f : \mathbb{R}^d \mapsto \mathbb{R}$ is twice differentiable at a point $\mathbf{x}_0$. Then

$$\begin{aligned} f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0) + \\ \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + o(\|\mathbf{x} - \mathbf{x}_0\|^2) \end{aligned}$$

Taylor's theorem: Lagrange form

- **Theorem (Taylor):** Suppose $f : \mathbb{R}^d \mapsto \mathbb{R}$ is differentiable on $N_r(\mathbf{x}_0)$. Then for any $\mathbf{x} \in N_r(\mathbf{x}_0)$, there exists $\bar{\mathbf{x}}$ on the line segment connecting $\mathbf{x}$ and $\mathbf{x}_0$ such that

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \mathbf{x}_0)$$

- **Theorem (Taylor):** Suppose $f : \mathbb{R}^d \mapsto \mathbb{R}$ is twice differentiable on $N_r(\mathbf{x}_0)$. Then for any $\mathbf{x} \in N_r(\mathbf{x}_0)$, there exists $\bar{\mathbf{x}}$ on the line segment connecting $\mathbf{x}$ and $\mathbf{x}_0$ such that

$$\begin{aligned} f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0) + \\ \tfrac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\bar{\mathbf{x}})(\mathbf{x} - \mathbf{x}_0) \end{aligned}$$

- "$\bar{\mathbf{x}}$ on the line segment connecting $\mathbf{x}$ and $\mathbf{x}_0$" means that there exists $w \in [0, 1]$ such that $\bar{\mathbf{x}} = w\mathbf{x} + (1 - w)\mathbf{x}_0$

Taylor's theorem: Third order

Theorem (Taylor): Suppose $f : \mathbb{R}^d \mapsto \mathbb{R}$ is three times differentiable on $N_r(\mathbf{x}_0)$. Then for any $\mathbf{x} \in N_r(\mathbf{x}_0)$, there exists $\bar{\mathbf{x}}$ on the line segment connecting $\mathbf{x}$ and $\mathbf{x}_0$ such that

$$\begin{aligned} f(\mathbf{x}) = f(\mathbf{x}_0) &+ \sum_{j=1}^{d} \frac{\partial f(\mathbf{x}_0)}{\partial x_j}(x_j - x_{0j}) \\ &+ \frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{d} \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_j \partial x_k}(x_j - x_{0j})(x_k - x_{0k}) \\ &+ \frac{1}{6}\sum_{j=1}^{d}\sum_{k=1}^{d}\sum_{\ell=1}^{d} \frac{\partial^3 f(\bar{\mathbf{x}})}{\partial x_j \partial x_k \partial x_\ell}(x_j - x_{0j})(x_k - x_{0k})(x_\ell - x_{0\ell}), \end{aligned}$$

where $\partial f(\mathbf{x}_0)/\partial x_j$ is shorthand for $\partial f(\mathbf{x})/\partial x_j$ evaluated at $\mathbf{x}_0$