Regression – Spectral Decomposition, Positive (Semi-)Definite Matrices, and the Statistical Linear Model

Class 07 (April 25)

Administrative Announcements

- **Upcoming break:** The course will pause for Passover. Lecture material will resume *after* the holiday; future meeting details will be communicated.
- **Linear Algebra Review:** A supplementary linear algebra review session will be offered this week with Mick via the Aviv platform. All are encouraged to attend, especially those feeling less comfortable with concepts such as diagonalization, spectral decomposition, and coordinate transformations.
- **Course Notes Clarification:** The present lecture revisits standard results from linear algebra, notably the diagonalization of symmetric matrices and properties of projection operators, as assumed background. While proofs are outlined for key results, full derivations may be omitted if previously seen in prerequisite courses; students are responsible for ensuring fluency with these results.
- **Homework Assignment:**
  - You are required to *find and write out the proof* that the eigenvalues of a projection matrix are 0 or 1, and that the rank equals the number of 1s among its eigenvalues.
  - Additionally, as an exercise, formally verify the claim that

    $$\mathrm{Cov}(Y_i, Y_j \mid X) = \mathrm{Cov}(\varepsilon_i, \varepsilon_j \mid X)$$

    under the model assumptions, for all i, j.
- **General Advice:** Students who have not recently reviewed key results from linear algebra—spectral decomposition, diagonalization, change of basis—are **strongly** encouraged to do so. Facility with coordinate systems and basis transitions is essential as we move forward in regression theory.

1 Motivation: The Power of Diagonalization

Why do we care about diagonalizing matrices, or about understanding the structure of symmetric or projection matrices?

In linear algebra, diagonalization allows us to "simplify" complicated linear transformations: when a matrix can be written as a diagonal in some basis, its action becomes transparently clear. Many statistical procedures—especially those involving projections, covariances, and regression—rely on understanding how matrices like these act, both abstractly and computationally.

We begin by revisiting core results from linear algebra before progressing to their implications for regression and the linear statistical model.

2 Spectral Decomposition and Diagonalization

2.1 Diagonalizable Matrices: Definitions and Meaning

Definition 2.1 (Diagonalizable Matrix)**.** A square matrix $A \in \mathbb{R}^{n\times n}$ is *diagonalizable* if there exists an invertible matrix P such that

$$A = PDP^{-1}$$

where D is a diagonal matrix.

Interpretation and Intuition:

Think of A as representing a linear operator on $\mathbb{R}^n$. If A is diagonalizable, then in an appropriate basis (specified by the columns of P), A acts by *scaling* each basis direction independently by the corresponding diagonal entry in D. All the intricate "mixing" seen in standard coordinates disappears—A simply stretches or shrinks along separate axes.

Explicitly: For a vector v, we can express the action of A:

$$\begin{aligned} \text{Let } P &= [p_1 \,|\, p_2 \,|\, \ldots \,|\, p_n] \text{ (columns as basis vectors)} \\ v &= P\tilde{v} \qquad \text{(express } v \text{ in basis } P) \\ Av &= PD\tilde{v}\,. \end{aligned}$$

Then, each component $\tilde{v}_i$ effectively gets multiplied by d_i, the ith diagonal entry of D.

2.2 Diagonalization of Symmetric Matrices (Spectral Theorem)

Theorem 2.2 (Spectral Theorem for Real Symmetric Matrices)**.** *Let $A \in \mathbb{R}^{n\times n}$ be symmetric, i.e., $A^T = A$. Then A is diagonalizable by an orthogonal matrix; i.e., there exists an orthogonal matrix U (so $U^T = U^{-1}$, columns are orthonormal) and a diagonal matrix D such that:*

$$A = UDU^T$$

where the diagonal entries of D are the (real) eigenvalues of A.

Remark 2.3. Diagonalization via an orthogonal matrix is much more robust numerically and conceptually: it says symmetric matrices are not only diagonalizable, but in a "nicest possible" way, with axes remaining orthogonal (no distortion).

Why is this useful? Because in regression and statistics, many important matrices (covariance matrices, projection matrices, etc.) are symmetric, and their properties—like positive-definiteness—relate directly to their eigenvalues.

3 Positive Definite and Positive Semi-Definite Matrices

3.1 Definitions and Key Properties

Definition 3.1 (Positive Definite and Positive Semi-Definite Matrices)**.** Let A be a symmetric matrix in $\mathbb{R}^{n\times n}$.

- A is called *positive definite* if for all $v \in \mathbb{R}^n$,
$$v^T A v > 0 \quad \text{unless} \quad v = 0.$$
- A is called *positive semi-definite* if for all $v \in \mathbb{R}^n$,
$$v^T A v \geq 0\,.$$

Here, $v^T A v$ is called a *quadratic form.*

Remark 3.2. Positivity is only defined for (real) symmetric matrices because otherwise $v^T A v$ may be complex or not have the right positivity properties across all v.

Theorem 3.3 (Eigenvalue Characterization)**.** *Let A be symmetric. Then,*

- *A is positive definite $\iff$ all its eigenvalues are strictly positive.*
- *A is positive semi-definite $\iff$ all its eigenvalues are non-negative.*

Proof Sketch. Given the spectral theorem, $A = UDU^T$, and for any v,
$$v^T A v = (U^T v)^T D (U^T v) = \sum_{i=1}^{n} \lambda_i (w_i)^2$$
where λ_i are the eigenvalues and w_i the components of $U^T v$. Clearly, if all $\lambda_i > 0$, then $v^T A v > 0$ unless $v = 0$; and similarly for the semi-definite condition. □

3.2 Square Roots of Positive Definite Matrices

A fundamental property is that every positive definite matrix has a unique positive definite square root, and (less uniquely) a symmetric matrix B with $B^2 = A$.

Theorem 3.4 (Matrix Square Root)**.** *If A is symmetric and positive definite, then there exists a symmetric matrix B such that $B^2 = A$. Moreover, if $A = UDU^T$, then $B = UD^{1/2}U^T$, where $D^{1/2}$ is the diagonal matrix whose entries are $\sqrt{d_i}$ (which are real and positive).*

Proof. Let $A = UDU^T$ as above. Define $B = UD^{1/2}U^T$. Then
$$B^2 = (UD^{1/2}U^T)(UD^{1/2}U^T) = UD^{1/2}(U^T U)D^{1/2}U^T = U(D^{1/2}D^{1/2})U^T = UDU^T = A$$
since $U^T U = I$ by orthogonality. □

Remark 3.5. For positive semi-definite A, similar constructions hold, but $D^{1/2}$ may have zeros.

4 Projection Matrices and Their Spectral Structure

4.1 Definition and Core Properties

Definition 4.1 (Projection Matrix)**.** A matrix $Q \in \mathbb{R}^{n\times n}$ is a projection matrix if

- $Q^2 = Q$ (idempotency)
- $Q^T = Q$ (symmetry)

Such matrices arise naturally when considering orthogonal projections onto subspaces, e.g., projecting a vector onto the column space of a matrix.

4.2 Spectral Decomposition of Projection Matrices

Theorem 4.2. *Suppose Q is a projection matrix of rank k (i.e., it projects onto a k-dimensional subspace). Then:*

- *Q is symmetric and idempotent ($Q^2 = Q$, $Q^T = Q$).*
- *Q is diagonalizable by an orthogonal matrix U: $Q = UDU^T$, where D is a diagonal matrix whose entries are either 0 or 1.*
- *Exactly k entries of D are 1; the rest (that is, $n-k$) are 0.*

Proof Sketch (Details as Homework). By the spectral theorem, any symmetric matrix is diagonalizable; the idempotency property restricts eigenvalues to satisfy $\lambda^2 = \lambda$, so $\lambda \in \{0, 1\}$. The rank equals the number of 1s among the eigenvalues. □

Example 4.3. Let Q be the matrix projecting $\mathbb{R}^3$ onto the xy-plane (i.e., $z = 0$). Then, in the standard basis, Q is

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

For any vector (x, y, z),

$$Q \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ y \\ 0 \end{pmatrix}$$

The eigenvalues are:

- 1 (with multiplicity 2): any vector in the xy-plane is left unchanged.
- 0 (multiplicity 1): vectors parallel to z are sent to 0.

This illustrates the general structure: projection matrices have eigenvalues 0 or 1.

Remark 4.4. More generally, in an appropriate orthogonal basis (usually given by the eigenvectors), any projection matrix appears as a diagonal matrix with k ones and $n-k$ zeros, possibly after a change of coordinates (rotation).

5 Transition to Regression: The Statistical Setting

5.1 From Deterministic Data to a Statistical Model

Traditionally, regression analysis begins with a set of observations, often arranged as

$$\begin{array}{ll} \text{Explanatory variables:} & x_{i1},\ x_{i2},\ \ldots, x_{ip} \quad (i = 1, \ldots, n) \\ \text{Outcome variable:} & y_i \quad (i = 1, \ldots, n) \end{array}$$

It's natural and productive to gather these into:

- a *design matrix* $X \in \mathbb{R}^{n \times (p+1)}$ (with an initial column of 1s for intercept),
- and a response vector $Y \in \mathbb{R}^n$.

$$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}, \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Remark 5.1. We usually require the columns of X to be linearly independent (i.e., $X^T X$ is invertible) for the least squares estimator to be unique.

So far, all the discussion has been *deterministic*: we have raw data, but **no probabilistic assumptions**.

5.2 Fitting the Least Squares Solution

The least squares estimator is

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

This minimizes the sum of squared residuals $\|Y - X\beta\|^2$ under the assumption that $X^T X$ is invertible.

Remark 5.2. Up to this point, our least squares solution is simply an algebraic result—no statistical model or stochastic assumption is needed.

5.3 Motivation for a Statistical Model

However, the ultimate goal of regression analysis is to make *inferences* about an underlying population—not just to describe a particular sample. That is, we wish to estimate how explanatory variables X relate to outcome Y in the larger population from which our data were drawn.

How can we formalize this?

We now **introduce a statistical model** in which each observed vector $(x_{i1}, \ldots, x_{ip}, y_i)$ is considered as a *realization* (i.e., outcome) of a random vector $(X_{i1}, \ldots, X_{ip}, Y_i)$ drawn from some joint distribution $\mathbb{P}$ over $\mathbb{R}^{p+1}$.

A Real-World Example (Rephrased from Class):

Suppose we sample a random person (say, a student) and record:

- Height: X_1
- Weight: X_2
- Gender: X_3 (categorical variable)

Every time we sample a person, we obtain a realization—a tuple of values drawn jointly from a population distribution $\mathbb{P}$ on $(X_1, X_2, X_3, \ldots)$. The assumption is that every sample is drawn "in the same way" (drawn from the same joint distribution).

6 The General Regression Decomposition

The statistical model allows us to clarify the aims of regression. At the highest generality:

Definition 6.1 (Conditional Expectation Decomposition)**.** Given a set of random variables $(X_{i1}, \ldots, X_{ip}, Y_i)$ drawn from a joint distribution $\mathbb{P}$, we can always write

$$Y_i = \mathbb{E}[Y_i \mid X_{i1}, \ldots, X_{ip}] + \big(Y_i - \mathbb{E}[Y_i \mid X_{i1}, \ldots, X_{ip}]\big)$$

or, to shorten notation,

$$Y_i = f(X_{i1}, \ldots, X_{ip}) + \varepsilon_i$$

where $f(X_{i1}, ..., X_{ip})$ is the **conditional mean** of Y_i given $X_{i1}, ..., X_{ip}$, and ε_i is the **residual (error) term**.

Key Property: By construction,

$$\mathbb{E}[\varepsilon_i \mid X_{i1}, \ldots, X_{ip}] = 0$$

That is, the residual is "centered" at zero given the values of the explanatory variables.

Remark 6.2. This decomposition always holds, whatever the (joint) distribution; the function $f(\cdots)$ may be highly nonlinear or complicated.

6.1 Interpretation and Purpose

- $f(\cdot)$ – the part of Y_i we *can* explain (given X). This is sometimes called the **systematic** or **signal** component.
- ε_i – the "noise" or "error" part, what Y does that is *not* explained by X.

Remark 6.3. Intuitively, if you know X, the best prediction for Y (in mean squared error) is the conditional expectation $f(X)$.

Example 6.4 (Motivating Example)**.** Suppose Y is the height of a person, and X is their weight. Then $f(X)$ is the average height of all people with weight X, and ε is the person's deviation from that average.

6.2 Key Mathematical Properties

Of particular importance:

$$\mathbb{E}[\varepsilon_i \mid X_{i1}, ..., X_{ip}] = 0\,, \qquad \mathbb{E}[\varepsilon_i] = 0$$

where the latter holds by the law of total expectation.

7 Assumptions of the Linear Statistical Model

Regression analysis via the linear model proceeds by making explicit, simplifying assumptions about the structure of $f(\cdot)$ and ε.

7.1 Linearity Assumption

Definition 7.1 (Linearity Assumption)**.** We *assume* that the conditional expectation of Y given X is a **linear function** of the predictors:

$$\mathbb{E}[Y_i \mid X_{i1}, ..., X_{ip}] = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$$

For notational convenience, we often write this as

$$\mathbb{E}[Y_i \mid X_{i0} = 1, X_{i1}, ..., X_{ip}] = \sum_{j=0}^{p} \beta_j X_{ij}$$

where $X_{i0} = 1$ for all i.

Remark 7.2. This is a **modeling assumption**: in general, the relation between Y and X may not be linear, but we often use linear approximation for interpretability and mathematical tractability.

7.2 Homoscedasticity & Uncorrelated Errors

Definition 7.3 (Equal Variance and Uncorrelated Errors)**.** The *homoscedastic (equal variance) linear model* further assumes:

$$\begin{aligned} \mathbb{E}[\varepsilon_i \mid X_{i1}, ..., X_{ip}] &= 0 \quad \text{(already discussed)} \\ \mathrm{Var}(\varepsilon_i \mid X_{i1}, ..., X_{ip}) &= \sigma^2 \quad \text{(does not depend on } X\text{)} \\ \mathrm{Cov}(\varepsilon_i, \varepsilon_j \mid X_{i1}, ..., X_{ip}, X_{j1}, ..., X_{jp}) &= 0 \qquad (i \neq j) \end{aligned}$$

Remark 7.4. Uncorrelated does not necessarily mean independent (though independence *implies* zero covariance).

Remark 7.5 (Relation to Observed Data)**.** Under these assumptions,

$$\mathrm{Var}(Y_i \mid X_{i1}, ..., X_{ip}) = \sigma^2$$

and likewise,

$$\mathrm{Cov}(Y_i, Y_j \mid X_{i1}, ..., X_{ip}, X_{j1}, ..., X_{jp}) = \mathrm{Cov}(\varepsilon_i, \varepsilon_j \mid X_{i1}, ..., X_{ip}, X_{j1}, ..., X_{jp}) = 0 \quad \text{for } i \neq j$$

Example 7.6 (Worked Example: Variance and Covariance in the Linear Model)**.** Let $Y_i = f(X_{i1}, ..., X_{ip}) + \varepsilon_i$, with model assumptions as above.

Then,

$$\mathrm{Var}(Y_i \mid X_{i1}, ..., X_{ip}) = \mathrm{Var}\left(f(X_{i1}, ..., X_{ip}) + \varepsilon_i \mid X_{i1}, ..., X_{ip}\right)$$

But since $f(\cdot)$ is deterministic given X, this is just

$$\mathrm{Var}(\varepsilon_i \mid X_{i1}, ..., X_{ip}) = \sigma^2$$

Similarly, for $i \neq j$,

$$\mathrm{Cov}(Y_i, Y_j \mid X) = \mathrm{Cov}(f(X_i), f(X_j)) + \mathrm{Cov}(\varepsilon_i, \varepsilon_j \mid X)$$

The first term is zero (they are constants), and the second is zero by assumption. Therefore,

$$\mathrm{Cov}(Y_i, Y_j \mid X) = 0$$

8 Rewriting the Model in Vector-Matrix Notation

The linear statistical model is often most conveniently expressed using vector and matrix notation, mirroring our earlier data arrangement.

8.1 Random Vectors and Matrices: Basic Definitions

Definition 8.1. A **random vector** Z is a finite collection of real random variables $(Z_1, ..., Z_n)$ defined on the same probability space.

Definition 8.2. A **random matrix** $Z \in \mathbb{R}^{n \times m}$ is a collection of real random variables Z_{ij}, $1 \leq i \leq n$, $1 \leq j \leq m$, all defined jointly.

Remark 8.3. In practice, arranging random variables into a matrix or as a long vector makes no probabilistic difference; it simply aids in aligning mathematical notation with later algebraic manipulations.

8.2 Expectation and Covariance Matrix

Definition 8.4 (Mean and Covariance)**.** The expectation of a random vector $Z = (Z_1, ..., Z_n)^T$ is the vector

$$\mathbb{E}[Z] = \begin{pmatrix} \mathbb{E}[Z_1] \\ \vdots \\ \mathbb{E}[Z_n] \end{pmatrix}$$

The (variance-)covariance matrix is the $n \times n$ matrix

$$\mathrm{Cov}(Z)_{ij} = \mathrm{Cov}(Z_i, Z_j) = \mathbb{E}[(Z_i - \mathbb{E}[Z_i])(Z_j - \mathbb{E}[Z_j])]$$

Remark 8.5. For a random matrix $Z \in \mathbb{R}^{n \times m}$, $\mathbb{E}[Z]$ is the $n \times m$ matrix of entrywise expectations $\mathbb{E}[Z_{ij}]$.

Example 8.6 (Variance and Covariance of Response Vector in Regression)**.** Let $Y = (Y_1, ..., Y_n)^T$ be the response vector under the linear model. Then,

$$\mathbb{E}[Y] = X\beta$$

where X is the design matrix (now treated as constant or fixed), and β is the parameter vector.

The covariance matrix is

$$\text{Cov}(Y \mid X) = \text{Cov}(\varepsilon \mid X) = \sigma^2 I_n$$

since the errors ε_i are homoscedastic and uncorrelated.

9 Summary and Looking Ahead

In this lecture, we've:

- Revisited the power of diagonalization, especially the spectral theorem for symmetric matrices.
- Explored projection matrices and proved all their eigenvalues are 0 or 1, tightly connecting to the geometry of regression.
- Transitioned from purely deterministic data to a statistical (random vector) model—the foundation for inference in regression.
- Laid out the central modeling assumptions of linear regression: linearity, homoscedasticity, and uncorrelated (or independent) errors.
- Recasted the model in modern vector-matrix notation, pausing for definitions of random vectors, random matrices, expectation, variance, and covariance matrices.

Next steps: We will use this formalism to analyze properties of the least squares estimator $\hat{\beta}$ in the statistical model, study its distribution, bias, variance, and confidence intervals.

Happy Passover! See you after the break.