Regression and Orthogonal Projections: Lecture Notes

Instructor: Amit

April 13, 2025

Course Announcements

- **Cameras on for Zoom lectures:** If you are comfortable, please turn on your cameras during classes; I would greatly appreciate it. If not, there is no obligation.
- **Upcoming Review:** There will be a linear algebra review session on Monday, at 10:00 AM, held over Zoom. This will be the only review for linear algebra in the course.
- **Independent Study:** From now until the end of the course, material will focus on statistics (especially multivariate statistical modeling) and its connections to linear algebra. Please take advantage of the term break to consolidate your understanding and fill in any gaps from previous tutorials, lessons, and exercises.
- **Practice and Review:** It is crucial that you can prove all core properties related to projection matrices as these are central to both the course and the final exam (as per last year's format).
- **Exercise Solutions:** Additional worked solutions and explanations will be provided soon. Please review them carefully and revisit concepts as needed.
- **Contact and Support:** Should you have any unresolved questions, please reach out. Your success is a top priority!

Contents

1 Introduction and Motivation

In modern statistical modeling and data analysis—regression in particular—geometry and linear algebra underlie everything. Understanding *projections*, especially onto subspaces spanned by the columns of a matrix, allows us to analyze how best to "approximate" data with models, how to decompose errors, and much, much more.

But why projection matrices? At its core, projection lets us answer questions such as:

> Given a large space of possible data (say, $\mathbb{R}^n$) and a subspace (for instance, all linear combinations of features in X), how do we find the *closest* point in this subspace to an arbitrary observation y?

Projection matrices provide the machinery to make this idea precise, to decompose vectors, and, as we'll see, to give beautiful geometric and statistical interpretations to regression.

2 The Geometry of Projection

2.1 Warm-Up: Projection in Lower Dimensions

Recall from high school or basic linear algebra: projecting a vector onto another vector in $\mathbb{R}^2$ or $\mathbb{R}^3$ is finding the closest point on a line to a given point.

Now imagine that instead of projecting onto a line, we're projecting onto a plane, or more generally, a subspace spanned by multiple vectors.

Key question: Given a vector $y \in \mathbb{R}^n$, and a subspace $V \subseteq \mathbb{R}^n$ (say, the column space of X), how do we find the point in V closest to y?

2.2 Projection onto Subspaces: The Basic Idea

Let X be an $n \times p$ matrix, with columns $x_1, x_2, \ldots, x_p$. Suppose $V = \mathrm{im}(X)$, the span of the columns.

Given any $y \in \mathbb{R}^n$, the *projection of y onto V* (denoted $P_V y$) is the vector in V which minimizes the distance to y:

$$P_V y = \operatorname*{argmin}_{v \in V} \|y - v\|_2.$$

Visualization: Draw the subspace V as a plane (when possible) in $\mathbb{R}^n$, y somewhere in space. The projection $P_V y$ is the "shadow" of y dropped perpendicularly onto V.

The error (or residual) is $e = y - P_V y$. Geometrically, e is orthogonal to V.

[Insert here, if possible, a diagram: a plane $V \subset \mathbb{R}^3$, a vector y above it, $P_V y$ as the foot of the perpendicular, e as the difference.]

Example 2.1 (Geometric Projection and Orthogonality). Suppose y is a point not in V. When we project y onto V, we are seeking the vector in V that is *closest* in Euclidean distance to y. By the Pythagorean Theorem (generalized), the difference e is at a right angle to all of V.

This geometric fact will yield powerful algebraic consequences!

3 Projection Matrices: Definition and Properties

3.1 Definition and Basic Properties

Definition 3.1 (Projection Matrix)**.** Let V be a subspace of $\mathbb{R}^n$, and let P be a linear operator on $\mathbb{R}^n$ (i.e., a matrix). P is called a *projection operator (matrix) onto V* if, for all y,

$$Py = \operatorname{argmin}_{v \in V} \|y - v\|_2.$$

P is called an *orthogonal projection* if the minimization uses the standard dot product.

A projection matrix enjoys two beautiful properties:

- **Idempotence:** $P^2 = P$. Projecting twice does nothing more than projecting once.
- **Symmetry:** $P^T = P$ (for the orthogonal case). Projection respects the inner product structure.

Remark 3.2. Idempotence is a geometric necessity: Once a vector is projected onto a subspace, applying the projection again does nothing—it's already in the subspace.

3.2 Construction: The Projection Matrix onto the Column Space of X

Suppose X is an $n \times p$ matrix with linearly independent columns.

Then the (orthogonal) projection matrix onto $\mathrm{im}(X)$ is:

$$P_X = X(X^T X)^{-1} X^T \tag{1}$$

Here, $X^T X$ is assumed invertible (i.e., X has full column rank).

Example 3.3 (Projection in the Simple Case). If X is $n \times 1$, i.e., a column vector x, then

$$P_X = \frac{xx^T}{x^T x}$$

This projects onto the line spanned by x.

Check for yourself: For $y \in \mathbb{R}^n$, $P_X y$ is the scalar multiple of x closest to y.

3.3 Key Properties of Orthogonal Projection Matrices

Let $P_X = X(X^T X)^{-1} X^T$.

- $P_X^2 = P_X$ (Idempotent)
- $P_X^T = P_X$ (Symmetric)
- $P_X X = X$ (Projection acts as identity on the image of X)
- If $w \in \mathrm{im}(X)$, then $P_X w = w$.
- If $v \in (\mathrm{im}(X))^\perp$ (orthogonal complement), then $P_X v = 0$, and v is not changed by projection onto $(\mathrm{im}(X))^\perp$.

3.4 Eigenvalues of Projection Matrices

Theorem 3.4 (Eigenvalues of Projection Matrices)**.** *Let P be an orthogonal projection matrix. Then all eigenvalues of P are either* 0 *or* 1*. The multiplicity of* 1 *equals* $\mathrm{rank}(P) = \dim(\mathrm{im}\, P)$*, and the multiplicity of* 0 *is* $n - \mathrm{rank}(P)$*.*

Proof Outline (with Insights). Recall that P is symmetric and idempotent.

Let u be an eigenvector: $Pu = \lambda u$. Applying P again (idempotence):

$$P^2 u = P(Pu) = P(\lambda u) = \lambda P u = \lambda^2 u$$

But also, $P^2 u = Pu = \lambda u$. Thus,

$$\lambda u = \lambda^2 u \implies (\lambda^2 - \lambda) u = 0$$

Since $u \neq 0$, either $\lambda = 0$ or $\lambda = 1$.

Multiplicity: The number of $\lambda = 1$ equals the rank of P (the dimension of the image), and the number of $\lambda = 0$ equals the dimension of the kernel. $\square$

4 The Spectral Perspective and Diagonalization

Orthogonal projection matrices, being real symmetric, can be diagonalized using an orthonormal basis of eigenvectors. That is:

$$P = U\Lambda U^T$$

where U is an orthogonal matrix whose columns are eigenvectors (basis), and Λ is diagonal with only 1s and 0s on the diagonal (positions corresponding to the subspace and its orthogonal complement).

Remark 4.1. This means that, in an appropriate basis, the action of the projection is simply to keep the coordinates associated with the subspace, and zero out the rest!

Example 4.2 (Geometric Interpretation). Imagine projecting onto a plane in $\mathbb{R}^3$. In a basis aligned with the plane and its orthogonal direction, the projection matrix simply leaves the in-plane components and zeros the perpendicular one.

5 Direct Sums, Orthogonal Complements, and Decomposition

5.1 Fundamental Theorem: Orthogonal Decomposition

Let V be a subspace of $\mathbb{R}^n$, and $V^\perp$ its orthogonal complement.

Theorem 5.1 (Direct Sum Decomposition)**.** *For every $y \in \mathbb{R}^n$ there exist unique $v \in V$ and $w \in V^\perp$ such that*

$$y = v + w$$

Moreover, $v = Py$ and $w = (I - P)y$, where P is the orthogonal projection onto V.

Sketch. The existence follows from Gram–Schmidt or via projection. Uniqueness by observing that $V \cap V^\perp = \{0\}$, since if z is in both, $z^T z = 0 \implies z = 0$. □

Example 5.2 (Regression Decomposition). In regression, for any y, writing

$$y = \hat{y} + e = P_X y + (I - P_X)y$$

gives a decomposition into the fitted value (*the projection onto the model space*) and the residual (*the error, orthogonal to the model space!*).

5.2 Image and Kernel Duality

A particularly elegant result is the following:

Theorem 5.3. *For any square matrix A,*

$$\operatorname{im}(A^T) = (\ker A)^\perp$$

That is, the image (column space) of A^T equals the orthogonal complement of the kernel of A.

Proof Outline. For any $w \in \operatorname{im}(A^T)$, $w = A^T k$ for some k. For $v \in \ker A$, $Av = 0$, so

$$v^T w = v^T A^T k = (Av)^T k = 0.$$

Therefore, w is orthogonal to all elements of $\ker A$. For the other direction, a similar argument applies. □

6 Statistical Modeling: Connecting Linear Algebra and Probability

6.1 Why Models Matter: Differentiating Math from Statistics

Mathematical and geometric observations about projections (e.g., the residual is orthogonal to the model space) are always true, regardless of where the data come from.

Statistical results, however, rely on assumptions about distributions (e.g., unbiasedness or optimality may depend on data being iid normal). It is crucial to distinguish between universally valid results and those which hinge on model assumptions.

6.2 Random Vectors, Expectations, and Covariances

Let z and w be random vectors of length n, with $z = (z_1, ..., z_n)^T$.

Definition 6.1 (Expectation of a Random Vector)**.** The expectation $\mathbb{E}[z]$ is the vector whose entries are $\mathbb{E}[z_i]$ individually.

Similarly, for a random matrix, expectation is entrywise.

Definition 6.2 (Covariance Matrix)**.** The covariance matrix of two random vectors $z, w \in \mathbb{R}^n$ is:

$$\mathrm{Cov}(z, w) = \mathbb{E}[(z - \mathbb{E}[z])(w - \mathbb{E}[w])^T]$$

If $z = w$, this is called the (variance-)covariance matrix of z.

- The diagonal entries are the variances of z_i.
- The off-diagonal entries are the covariances between z_i and z_j.
- The covariance matrix is always symmetric: $\mathrm{Cov}(z_i, z_j) = \mathrm{Cov}(z_j, z_i)$.

Theorem 6.3 (Linearity and Transformation)**.** *Let A, B be deterministic matrices and z a random vector. Then:*

$$\mathrm{Cov}(Az, Bz) = A\,\mathrm{Cov}(z)B^T$$

This is essential when analyzing the variability of linear combinations of random vectors.

Remark 6.4. For a deterministic row vector a,

$$\mathrm{Var}(a^T z) = a^T\,\mathrm{Cov}(z)a$$

This gives the variance of any linear combination of the entries of z.

6.3 Worked Example: Bernoulli Random Vectors

Example 6.5 (Covariances in a Simple Bernoulli Model). Let $z = (x, y, m)$, where:

- x is Bernoulli with parameter p,
- y is Bernoulli with parameter q,

- $m = xy$.

What is the support, expectation, and covariance matrix?

<u>Support:</u>

$$\text{supp} = \{0,1\} \times \{0,1\} \times \{0,1\}$$

<u>Expectation:</u>

$$\mathbb{E}[x] = p, \quad \mathbb{E}[y] = q, \quad \mathbb{E}[m] = \mathbb{E}[xy] = pq$$

<u>Covariance Matrix:</u> Let us compute each entry.

$$\begin{aligned}
\text{Var}(x) &= p(1-p) \\
\text{Cov}(x,y) &= \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] = pq - pq = 0 \\
\text{Cov}(x,m) &= \mathbb{E}[xm] - \mathbb{E}[x]\mathbb{E}[m] = \mathbb{E}[x^2y] - p(pq) \\
&= \mathbb{E}[xy] - p^2q = pq - p^2q = pq(1-p) \\
\text{Var}(y) &= q(1-q) \\
\text{Cov}(y,m) &= \mathbb{E}[ym] - \mathbb{E}[y]\mathbb{E}[m] = \mathbb{E}[xy^2] - q(pq) \\
&= \mathbb{E}[xy] - pq^2 = pq - pq^2 = pq(1-q) \\
\text{Var}(m) &= pq(1-pq)
\end{aligned}$$

Hence the covariance matrix is:

$$\text{Cov}(z) = \begin{pmatrix} p(1-p) & 0 & pq(1-p) \\ 0 & q(1-q) & pq(1-q) \\ pq(1-p) & pq(1-q) & pq(1-pq) \end{pmatrix}$$

Remark 6.6. Notice how dependent variables (like $m = xy$) create interesting covariance patterns even if x and y themselves are independent.

7 The Linear Statistical Model

Let us now rigorously introduce the core statistical model underpinning regression and projections.

Definition 7.1 (Linear Model)**.** We observe $y \in \mathbb{R}^n$ (*response vector*), an $n \times p$ matrix X of predictors, and seek to model y as a linear combination:

$$y = X\beta + \varepsilon$$

where:

- X is the *design matrix*, entries fixed (can be random in other contexts).
- $\beta \in \mathbb{R}^p$ is the unknown vector of coefficients/parameters.
- $\varepsilon \in \mathbb{R}^n$ is the vector of errors/noise.

We typically assume:

- $\mathbb{E}[\varepsilon|X] = 0$
- $\mathrm{Cov}(\varepsilon_i, \varepsilon_j|X) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases}$

The errors are uncorrelated with equal variance (*homoskedasticity*).

Why is this useful? This model supports the inference and estimation of the "true" relationship between predictors X and the response y, acknowledging randomness/noise.

Example 7.2 (Regression Setup). Suppose we are predicting salary y based on education, parents' income, and number of children. Each variable is a column in X, and β encodes the effect size of each variable. The error vector ε absorbs all variation not explained linearly by these predictors.

Remark 7.3 (Implications of Model Assumptions). If the errors ε are correlated, or have non-constant variance, the effective information in the data is reduced (some observations are redundant), invalidating standard inferential procedures.

8 Properties and Interpretation of Projection in the Regression Model

8.1 Fitted Values and Residuals

Given the linear model $y = X\beta + \varepsilon$ and the projection matrix P_X,

- The vector of fitted values is $\hat{y} = P_X y = X(X^T X)^{-1} X^T y$.
- The vector of residuals is $e = y - \hat{y} = (I - P_X)y$.
- By construction, e is orthogonal to $\mathrm{im}(X)$.

8.2 Summary of Key Points

- The geometry of projection is at the heart of regression theory.
- Projection matrices decompose data vectors into components explained by the model and pure error (residual).
- Statistical properties (like variance, unbiasedness) depend on additional assumptions about distributions.
- Full understanding of projections involves skill in both linear algebra and statistical reasoning.

9 Further Directions and Preparation

Next Steps and Recommendations

- **Review core properties and proofs**: Being able to prove the key results around projections and their application to regression is essential.
- **Revisit previous exercises and solutions**, especially examples involving projections, decomposition, and interpretation of statistical models.
- **Prepare for the next session:** We will deepen the statistical modeling (including random vectors, distributions, and the connection to regression) and will discuss the assumptions underlying statistical inference in linear contexts.
- **Administrative follow-up:** View all updated practice solutions and plan accordingly for (i) completing open assignments, (ii) course requirements, and (iii) lecture participation.

Thank you, and have an enjoyable and productive break!