Projections in Finite Dimensions & Least Squares

BST 257: Theory and Methods for Causality II
Alex Levis, Fall 2021

1 Projections in Hilbert spaces

A (real) Hilbert space is a vector space H equipped with an inner product $\langle \cdot , \cdot \rangle : H \times H \to \mathbb{R}$, such that H is complete with respect to the norm induced by its inner product. Two examples we saw were the finite-dimesional Euclidian space $\mathbb{R}^k$, for any $k \in \mathbb{N}$, with the inner product being the standard dot product, and the space of finite variance real-valued functions of $\boldsymbol{X} \sim P$, denoted $L_2(P)$, with inner product $\langle g, h\rangle = \mathbb{E}_P(g(\boldsymbol{X})h(\boldsymbol{X}))$.

A fundamental fact about Hilbert spaces is that for any closed linear subspace $U \subseteq H$, projections onto U exist and are unique. This concept can be defined in two equivalent ways: let $v \in H$, then the projection of v onto U, denoted $\Pi_U(v)$ for now, is the unique vector satisfying:

(1) $\Pi_U(v) \in U$, and

(2) $v - \Pi_U(v) \perp U$.

Equivalently, (1) and (2) hold if and only if

$$\Pi_U(v) = \underset{m \in U}{\arg\min} \|v - m\|,$$

which is to say $\Pi_U(v)$ is the closest vector in U to v. Essential properties of the projection operator $\Pi_U : H \to U$, that follow from the definition, are that it is

(a) Linear: $\Pi_U(av_1 + bv_2) = a\Pi_U(v_1) + b\Pi_U(v_2)$, for all $a, b \in \mathbb{R}$, $v_1, v_2 \in H$.

(b) Self-adjoint: $\langle \Pi_U(v_1), v_2\rangle = \langle v_1, \Pi_U(v_2)\rangle$, for all $v_1, v_2 \in H$.

(c) Idempotent: $\Pi_U(\Pi_U(v)) = \Pi_U(v)$, for all $v_1, v_2 \in H$.

A convenient fact about finite-dimensional linear subspaces of a Hilbert space is that they are always closed. The theory of least squares and linear models can then flow nicely by thinking about projections in certain vector spaces. Let's explore this next.

2 Population least squares

Consider the finite-dimensional subspace of H spanned by $v_1, \ldots, v_k \in H$:

$$V_0 = \mathscr{L}(v_1, \ldots, v_k) := \left\{ \sum_{j=1}^{k} \alpha_j v_j \,\middle|\, \alpha_1, \ldots, \alpha_k \in \mathbb{R} \right\}.$$

For any $v \in H$, there exists $\boldsymbol{\alpha} = [\alpha_1 \cdots \alpha_k]^T \in \mathbb{R}^k$ such that $\Pi_{V_0}(v) = \sum_{j=1}^k \alpha_j v_j$, by definition of projection. It can be shown that $\boldsymbol{\alpha}$ is a solution to the so-called *normal equations*,

$$\boldsymbol{M}\boldsymbol{\alpha} = \boldsymbol{\nu}, \tag{1}$$

where the Gram matrix $\boldsymbol{M} \in \mathbb{R}^{k \times k}$ is given by $[\boldsymbol{M}]_{i,j} = \langle v_i, v_j \rangle$, and $\boldsymbol{\nu} = [\langle v, v_1 \rangle \cdots \langle v, v_k \rangle]^T \in \mathbb{R}^k$. Moreover,

- There always exists a solution to (1).
- There is a unique solution if and only if $\operatorname{rank}(\boldsymbol{M}) = k$.
- There is a unique solution if and only if $\{v_1, \dots, v_k\}$ are linearly independent.

In the case where the solution is unique, the matrix $\boldsymbol{M}$ has an inverse, and we have a formula for *the* solution to (1), given by

$$\boldsymbol{\alpha} = \boldsymbol{M}^{-1}\boldsymbol{\nu}.$$

Consider the typical regression setting, in which observe a sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ of iid copies of $(\boldsymbol{X}, Y) \sim P$, where $\boldsymbol{X} = (X_1, \dots, X_k)^T \in \mathbb{R}^k$. Assuming the k components of $\boldsymbol{X}$ have finite variance, we can consider projection of $Y \in L_2(P)$ onto $\mathcal{X} = \mathscr{L}(X_1, \dots, X_k) \subseteq L_2(P)$. Since this is a k-dimensional subspace, we know by (1) that there exists $\boldsymbol{\beta}^\dagger \in \mathbb{R}^k$ such that

$$\mathbb{E}_P[\mathbf{X}\mathbf{X}^T]\boldsymbol{\beta}^\dagger = \mathbb{E}_P[\mathbf{X}Y] \iff \boldsymbol{X}^T\boldsymbol{\beta}^\dagger = \Pi_{\mathcal{X}}(Y).$$

If the Gram matrix $\mathbb{E}_P[\boldsymbol{X}\boldsymbol{X}^T]$ has full rank, which we shall assume, then there is only one solution to the normal equations:

$$\boldsymbol{\beta}(P) := \mathbb{E}_P[\boldsymbol{X}\boldsymbol{X}^T]^{-1}\mathbb{E}_P[\boldsymbol{X}Y], \text{ so that } \Pi_{\mathcal{X}}(Y) = \boldsymbol{X}^T\boldsymbol{\beta}(P) = \mathbb{E}_P[Y\boldsymbol{X}^T]\mathbb{E}_P[\boldsymbol{X}\boldsymbol{X}^T]^{-1}\boldsymbol{X}.$$

Meanwhile, by the equivalent characterization of projections,

$$\boldsymbol{\beta}(P) = \underset{\boldsymbol{\beta} \in \mathbb{R}^k}{\arg\min} \|Y - \boldsymbol{\beta}^T\boldsymbol{X}\|_{L_2(P)}.$$

Thus, $\boldsymbol{\beta}(P)$ can be interpreted as the population least squares parameter, as

$$\|Y - \boldsymbol{\beta}^T\boldsymbol{X}\|_{L_2(P)}^2 = \mathbb{E}_P[(Y - \boldsymbol{\beta}^T\boldsymbol{X})^2].$$

Everything we have done thus far is completely agnostic to a statistical model for P (other than second moment restrictions). So how does this relate to linear models? Suppose we assert the linear model

$$\mathbb{E}_P(Y \mid \boldsymbol{X}) = \boldsymbol{X}^T\boldsymbol{\beta}_0.$$

A classic result is that the conditional expectation of Y given $\boldsymbol{X}$ also has a projection interpretation:

$$\mathbb{E}_P(Y \mid \boldsymbol{X}) = \underset{g}{\arg\min} \|Y - g(\boldsymbol{X})\|_{L_2(P)}^2.$$

This implies that $\boldsymbol{\beta}_0 = \boldsymbol{\beta}(P)$ — can you see this? But interestingly, the parameter $\boldsymbol{\beta}(P)$ is well-defined regardless.

In class, we use the notation $\Pi[Y \mid \boldsymbol{X}]$ in lieu of $\Pi_{\mathcal{X}}(Y)$. Although slightly more ambiguous, this notation is convenient, and we overload it to also mean the stacked projections when the input is a vector: if $\boldsymbol{W} = (W_1, \ldots, W_p)^T \in L_2(P)^p$, then

$$\Pi[\boldsymbol{W} \mid \boldsymbol{X}] \coloneqq \begin{bmatrix} \Pi_{\mathcal{X}}(W_1) \\ \vdots \\ \Pi_{\mathcal{X}}(W_p) \end{bmatrix} = \mathbb{E}_P[\boldsymbol{W}\boldsymbol{X}^T]\mathbb{E}_P[\boldsymbol{X}\boldsymbol{X}^T]^{-1}\boldsymbol{X}.$$

By extension of univariate projection properties, $\Pi[\,\cdot \mid \boldsymbol{X}]$ is a linear function of its input, has components belonging to $\mathcal{X}$, and satisfies the residual orthogonality requirement

$$\begin{aligned}
& W_\ell - \Pi[W_\ell \mid \boldsymbol{X}] \perp \mathcal{X}, \text{ for } \ell = 1, \ldots, p, \\
\iff & 0 = \mathbb{E}_P\left(\{W_\ell - \Pi[W_\ell \mid \boldsymbol{X}]\}\, \boldsymbol{X}^T\boldsymbol{\alpha}\right), \text{ for any } \boldsymbol{\alpha} \in \mathbb{R}^k, \text{ for } \ell = 1, \ldots, p \\
\iff & \mathbf{0}_p = \mathbb{E}_P\left(\{\boldsymbol{W} - \Pi[\boldsymbol{W} \mid \boldsymbol{X}]\}\, \boldsymbol{X}^T\boldsymbol{\alpha}\right), \text{ for any } \boldsymbol{\alpha} \in \mathbb{R}^k, \\
\iff & \mathbf{0}_{p\times k} = \mathbb{E}_P\left(\{\boldsymbol{W} - \Pi[\boldsymbol{W} \mid \boldsymbol{X}]\}\, \boldsymbol{X}^T\right).
\end{aligned}$$

Inspecting this requirement, we can prove property 1 of Π from page 49 of the notes: if $\boldsymbol{X}$ includes a non-zero constant, say $X_1 \equiv 1$, then $\mathbb{E}_P(\boldsymbol{W} - \Pi[\boldsymbol{W} \mid \boldsymbol{X}]) = \mathbf{0}_p$ — just look at the first column of the equation on the last line.

3 Sample least squares

Everything stated in the last section can be said for any probability distribution P such that $X_1, \ldots, X_p, Y \in L_2(P)$. Consider the empirical law, P_n let $\mathbb{P}_n \equiv \mathbb{E}_{P_n}$ denote expectation under P_n, and let $\Pi_{\mathcal{X}}^{(n)}$ denote projection onto $\mathcal{X} = \mathscr{L}(X_1, \ldots, X_k) \subseteq L_2(P_n)$. Assuming the Gram matrix $\mathbb{P}_n[\boldsymbol{X}\boldsymbol{X}^T] = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T$ has (full) rank k,

$$\boldsymbol{\beta}(P_n) = \mathbb{P}_n[\boldsymbol{X}\boldsymbol{X}^T]^{-1}\mathbb{P}_n[\boldsymbol{X}Y], \text{ so that } \Pi_{\mathcal{X}}^{(n)}(Y) = \boldsymbol{X}^T\boldsymbol{\beta}(P_n) = \mathbb{P}_n[Y\boldsymbol{X}^T]\mathbb{P}_n[\boldsymbol{X}\boldsymbol{X}^T]^{-1}\boldsymbol{X}.$$

It turns out that $\boldsymbol{\beta}(P_n)$ is exactly the same as the sample least squares estimate $\widehat{\boldsymbol{\beta}}$. This can be seen readily from

$$\boldsymbol{\beta}(P_n) = \underset{\boldsymbol{\beta}\in\mathbb{R}^k}{\arg\min} \|Y - \boldsymbol{\beta}^T\boldsymbol{X}\|_{L_2(P_n)},$$

since

$$\|Y - \boldsymbol{\beta}^T\boldsymbol{X}\|_{L_2(P_n)}^2 = \mathbb{P}_n[(Y - \boldsymbol{\beta}^T\boldsymbol{X})^2] = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \boldsymbol{\beta}^T\boldsymbol{X}_i\right)^2$$

is the least squares criterion, minimized by $\widehat{\boldsymbol{\beta}}$.

Finally, note that in class, we used the notation $\Pi_n[Y \mid \boldsymbol{X}]$ instead of $\Pi_{\mathcal{X}}^{(n)}(Y)$, and we can also extend to arbitrary $\boldsymbol{W} = (W_1, \ldots, W_p)^T \in L_2(P_n)^p$:

$$\Pi_n[\boldsymbol{W} \mid \boldsymbol{X}] \coloneqq \begin{bmatrix} \Pi_{\mathcal{X}}^{(n)}(W_1) \\ \vdots \\ \Pi_{\mathcal{X}}^{(n)}(W_p) \end{bmatrix} = \mathbb{P}_n[\boldsymbol{W}\boldsymbol{X}^T]\mathbb{P}_n[\boldsymbol{X}\boldsymbol{X}^T]^{-1}\boldsymbol{X}.$$

As before, if $\boldsymbol{X}$ includes a non-zero constant under P_n, say $X_{i,1} = 1$ for $i = 1, \ldots, n$, then

$$\mathbb{P}_n(\boldsymbol{W} - \Pi_n[\boldsymbol{W} \mid \boldsymbol{X}]) = \mathbf{0}_p.$$

3.1 Properties of $\widehat{\beta}$

It is not too hard to see that $\widehat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}(P)$: by the weak law of large numbers

$$\mathbb{P}_n[\boldsymbol{X}\boldsymbol{X}^T] \xrightarrow{P} \mathbb{E}_P(\boldsymbol{X}\boldsymbol{X}^T) \text{ and } \mathbb{P}_n[\boldsymbol{X}Y] \xrightarrow{P} \mathbb{E}_P[\boldsymbol{X}Y].$$

Thus, by the continuous mapping theorem — in reality, this requires some care, as there may be some probability that $\operatorname{rank}(\mathbb{P}_n[\boldsymbol{X}\boldsymbol{X}^T]) < k$ — and limit laws for convergence in probability,

$$\widehat{\boldsymbol{\beta}} = \mathbb{P}_n[\boldsymbol{X}\boldsymbol{X}^T]^{-1}\mathbb{P}_n[\boldsymbol{X}Y] \xrightarrow{P} \mathbb{E}_P[\boldsymbol{X}\boldsymbol{X}^T]^{-1}\mathbb{E}_P[\boldsymbol{X}Y] = \boldsymbol{\beta}(P).$$

With some more work, we can establish asymptotic normality of $\widehat{\boldsymbol{\beta}}$:

$$\begin{aligned}
\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}(P)\right) &= \sqrt{n}\left(\mathbb{P}_n[\boldsymbol{X}\boldsymbol{X}^T]^{-1}\mathbb{P}_n[\boldsymbol{X}Y] - \boldsymbol{\beta}(P)\right) \\
&= \sqrt{n}\mathbb{P}_n[\boldsymbol{X}\boldsymbol{X}^T]^{-1}\left(\mathbb{P}_n[\boldsymbol{X}Y] - \mathbb{P}_n[\boldsymbol{X}\boldsymbol{X}^T]\boldsymbol{\beta}(P)\right) \\
&= \sqrt{n}\mathbb{P}_n[\boldsymbol{X}\boldsymbol{X}^T]^{-1}\mathbb{P}_n\left[\boldsymbol{X}\left(Y - \boldsymbol{X}^T\boldsymbol{\beta}(P)\right)\right]
\end{aligned}$$

Note that $\mathbb{E}_P[\boldsymbol{X}(Y - \boldsymbol{X}^T\boldsymbol{\beta}(P))] = \mathbb{E}_P[(Y - \Pi[Y \mid \boldsymbol{X}])\boldsymbol{X}^T]^T = \mathbf{0}_k$, by orthogonality, so we may use the central limit theorem to deduce that

$$\sqrt{n}\mathbb{P}_n\left[\boldsymbol{X}\left(Y - \boldsymbol{X}^T\boldsymbol{\beta}(P)\right)\right] \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}_k, \boldsymbol{A}),$$

where $\boldsymbol{A} = \operatorname{Var}_P(\boldsymbol{X}\left(Y - \boldsymbol{X}^T\boldsymbol{\beta}(P)\right) \in \mathbb{R}^{k\times k}$. Letting $\boldsymbol{B} = \mathbb{E}_P(\boldsymbol{X}\boldsymbol{X}^T) \in \mathbb{R}^{k\times k}$ and applying a multivariate version of Slutsky's theorem,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}(P)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}_k, \boldsymbol{V}(P)),$$

where $\boldsymbol{V}(P) = \boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{B}^{-1}$.

One last reminder: all of this *does not rely on an underlying linear model* for $\mathbb{E}_P(Y \mid \boldsymbol{X})$.