# Projections and Random Vectors (Lab 3)

BST 235: Advanced Regression and Statistical Learning
Alex Levis, Fall 2019

## 1 Linear algebra review

Let's recap what we have accomplished thus far in our exploration of linear algebra.

(1) We first defined abstract vector spaces and linear subspaces, then covered basis and dimension.

(2) Last lab we defined linear maps, discussed their invertibility, introduced coordinates, and argued that for finite-dimensional vector spaces, linear maps can be represented by a matrix. As one special case, we proved that every linear map $T : \mathbb{R}^d \to \mathbb{R}^n$ can be associated with a (unique) matrix $M_T \in \mathbb{R}^{n\times d}$ such that $T(\mathbf{x}) = M_T\mathbf{x}$, for all $\mathbf{x} \in \mathbb{R}^d$.

(3) In lecture, we introduced inner product spaces (i.e., vector spaces with an associated inner product), defined orthogonality, orthogonal complements, direct orthogonal sums, and discussed projections onto finite-dimensional linear subspaces.

Since we have not yet talked about inner product spaces in lab, we now elaborate the results we have developed for point (3) above.

## 2 The projection operator

Let $(V, \langle\cdot,\cdot\rangle)$ be a real inner product space, $V_0 \subseteq V$ a finite-dimensional linear subspace of $V$. In several steps, we proved that for every $v \in V$, we could associate a unique vector $P_{V_0}(v) \in V_0$, called the projection of $v$ onto $V_0$, satisfying one of the following two equivalent criteria:

- $v - P_{V_0}(v) \perp V_0 \iff v - P_{V_0}(v) \in V_0^{\perp} \iff \langle v - P_{V_0}(v), w\rangle = 0$ for all $w \in V_0$,
- $P_{V_0}(v) = \arg\min_{w\in V_0}\|v - w\|$.

At the end of lecture 5, we showed that the function $P_{V_0} : V \to V_0$ held several nice properties. Specifically, we proved that $P_{V_0}$ is

(i) Linear: $P_{V_0} \in \mathcal{L}(V, V_0)$,

(ii) Idempotent: $P_{V_0} \circ P_{V_0} = P_{V_0}$,

(iii) Self-adjoint: $\langle P_{V_0}(v_1), v_2\rangle = \langle v_1, P_{V_0}(v_2)\rangle$, for all $v_1, v_2 \in V$.

While properties (ii) and (iii) will turn out to be very useful, property (i) is perhaps even more important, as it allows us to associate $P_{V_0}$ with a *projection matrix*! Letting $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(d)} \in \mathbb{R}^n$ denote the columns of the design matrix, we can set $V_0 = \mathcal{C}(\mathbb{X}) = \mathscr{L}(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(d)})$, and consider the matrix corresponding to the linear map $P_{\mathcal{C}(\mathbb{X})} : \mathbb{R}^n \to \mathbb{R}^n$,

$$\widehat{P}_{\mathbb{X}} := M_{P_{\mathcal{C}(\mathbb{X})}} \in \mathbb{R}^{n\times n},$$

the famous *hat matrix*, i.e., the projection matrix for the linear subspace $\mathcal{C}(\mathbb{X})$.

We will see that in the linear model, a least squares estimator $\widehat{\boldsymbol{\beta}}$ must satisfy

$$\mathbb{X}\widehat{\boldsymbol{\beta}} = \widehat{P}_{\mathbb{X}}\mathbf{Y} = P_{\mathcal{C}(\mathbb{X})}(\mathbf{Y}).$$

In words, this says that the fitted values are the projection of the outcome onto the space spanned by the columns of the design matrix. As an exercise, given what we learned last lab, think about what properties (ii) and (iii) of projection imply for the hat matrix $\widehat{P}_{\mathbb{X}}$.

**Exercise 1.** Given vector space $V$, and a list of linearly independent vectors $\{v_1, \dots, v_k\} \in V$, we have seen the Gram-Schmidt procedure that produces $\{v_1^*, \dots, v_k^*\}$, an orthogonal basis for $\mathscr{L}(v_1, \dots, v_k)$. Show that the procedure "fails" if $\{v_1, \dots, v_k\}$ are linearly dependent.

By Exercise 5 in Lab 1, we know that linear dependence implies that $\exists\, j \in \{1, \dots, k\}$, and $\alpha_1, \dots, \alpha_{j-1} \in \mathbb{F}$, such that $v_j = \sum_{\ell=1}^{j-1} \alpha_\ell v_\ell$. Without loss of generality, let $j$ be the minimal such index, so that $\{v_1, \dots, v_{j-1}\}$ are linearly independent. If $j = 1$, then $v_1 = 0_V$ and Gram-Schmidt fails immediately, as $v_1^* := v_1 = 0_V$ means the resulting set of vectors cannot be non-zero and orthogonal. Otherwise, at the $j$-th step, we will compute

$$v_j^* = v_j - p(v_j \,|\, \mathscr{L}(v_1^*, \dots, v_{j-1}^*)) = v_j - v_j = 0_V,$$

since $v_j \in \mathscr{L}(v_1, \dots, v_{j-1}) = \mathscr{L}(v_1^*, \dots, v_{j-1}^*)$. Again, this fails to produce an orthogonal basis.

**Exercise 2.** Let $(V, \langle \cdot, \cdot \rangle)$ be a real inner product space, and let $V_0 \subseteq V$ be a finite-dimensional linear subspace of $V$. Show that

$$\|P_{V_0}(v)\| \leq \|v\|, \text{ for all } v \in V.$$

In other words, projection is a contraction mapping. If you have seen operator norms before, this is equivalent to the statement that $\|P_{V_0}\|_{\text{op}} \leq 1$. Recall in the regression setting that $\mathbb{E}_P(Y \,|\, \mathbf{X}) = \arg\min_g (Y - g(\mathbf{X}))^2$. Considering the (complete) inner product space

$$L_2(\mathbf{X}, Y) = \{g(\mathbf{X}, Y) \,|\, g \text{ measurable}, \mathbb{E}(g(\mathbf{X}, Y)^2) < \infty\},$$

its (closed) linear subspace $L_2(\mathbf{X}) = \{g(\mathbf{X}) \,|\, g \text{ measurable}, \mathbb{E}(g(\mathbf{X})^2) < \infty\}$, with

$$\langle g_1, g_2 \rangle = \mathbb{E}_P(g_1(\mathbf{X}, Y) g_2(\mathbf{X}, Y)),$$

what does this result say about the effect of conditional expectation on $\mathbf{X}$? What if $\mathbb{E}_P(Y) = 0$? (Note that projection in this infinite-dimensional setting works just as we have learned!)

For any $v \in V$,

$$\|v\|^2 = \|v - P_{V_0}(v)\|^2 + \|P_{V_0}(v)\|^2 \geq \|P_{V_0}(v)\|^2,$$

by the Pythagorean theorem, as $v - P_{V_0}(v) \perp P_{V_0}(v)$. In the conditional expectation setting, noting that $\mathbb{E}(Y \,|\, \mathbf{X}) = P_{L_2(\mathbf{X})}(Y)$, this result translates to

$$\mathbb{E}[\mathbb{E}(Y \,|\, \mathbf{X})^2] \leq \mathbb{E}[Y^2].$$

If actually $\mathbb{E}(Y) = 0$, then equivalently

$$\text{Var}(\mathbb{E}(Y \,|\, \mathbf{X})) \leq \text{Var}(Y),$$

so conditional expectation decreases variance.

**Exercise 3.** Let $(V, \langle\cdot,\cdot\rangle)$ be a real inner product space, and let $V_1 \subseteq V_2$, where $V_1$ and $V_2$ are finite-dimensional linear subspaces of $V$. Show that

$$P_{V_1} = P_{V_1} \circ P_{V_2} = P_{V_2} \circ P_{V_1},$$

and think about what this would mean for the corresponding projection matrices.

The result $P_{V_2} \circ P_{V_1} = P_{V_1}$, is immediate, as for any $v \in V$, $P_{V_1}(v) \in V_1 \subseteq V_2$, so further projection onto $V_2$ does nothing. For the other equality, let $v \in V$ be arbitrary and we will use the definition. First $P_{V_1}(P_{V_2}(v)) \in V_1$, by definition of $P_{V_1}$. Second, for any $w \in V_1$,

$$\begin{aligned}\langle v - P_{V_1}(P_{V_2}(v)), w\rangle &= \langle v, w\rangle - \langle P_{V_1}(P_{V_2}(v)), w\rangle \\ &= \langle v, w\rangle - \langle v, P_{V_2}(P_{V_1}(w))\rangle \\ &= \langle v, w\rangle - \langle v, w\rangle = 0\end{aligned}$$

since projection is self-adjoint and $w \in V_1 \subseteq V_2$. Therefore, $v - P_{V_1}(P_{V_2}(v)) \perp V_1$, and we know $P_{V_1}(P_{V_2}(v)) = P_1(v)$. In terms of projection matrices, this says that if $V_1 \subseteq V_2$,

$$\widehat{P}_{V_1} = \widehat{P}_{V_1}\widehat{P}_{V_2} = \widehat{P}_{V_2}\widehat{P}_{V_1},$$

given what we know about linear map composition and matrix multiplication.

## 3 Random vectors and matrices

A random vector is simply an ordered collection of random variables. Abstractly, on probability space $(\Omega, \mathcal{A}, P)$, the vector-valued function $\mathbf{X} : \Omega \to \mathbb{R}^k$ is called a *random vector* if and only if $X_1, \ldots, X_k : \Omega \to \mathbb{R}$ are random variables (i.e., Borel measurable), where $\mathbf{X} = [X_1 \cdots X_k]^T$. The expectation of $\mathbf{X}$ is defined to be the vector of expected values of each of its components:

$$\mathbb{E}(\mathbf{X}) = \begin{bmatrix}\mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_k)\end{bmatrix} \in \mathbb{R}^k,$$

whenever all the expectations exist. Analogously, $\mathbb{M} : \Omega \to \mathbb{R}^{m\times n}$ is a *random matrix* if each of its elements is a random variable, the its expectation is the matrix of the element-wise expected valued, when they exist. The cross-covariance between two random vectors $\mathbf{X} \in \mathbb{R}^k$, and $\mathbf{Y} \in \mathbb{R}^\ell$ is the $(k \times \ell)$ matrix given by

$$\operatorname{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{Y} - \mathbb{E}(\mathbf{Y})^T)\right] = \begin{bmatrix}\operatorname{Cov}(X_1, Y_1) & \operatorname{Cov}(X_1, Y_2) & \cdots & \operatorname{Cov}(X_1, Y_\ell) \\ \operatorname{Cov}(X_2, Y_1) & \operatorname{Cov}(X_2, Y_2) & \cdots & \operatorname{Cov}(X_2, Y_\ell) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_k, Y_1) & \operatorname{Cov}(X_k, Y_2) & \cdots & \operatorname{Cov}(X_k, Y_\ell)\end{bmatrix},$$

and the covariance matrix for the random vector $\mathbf{X}$ is the $(k \times k)$ matrix

$$\operatorname{Cov}(\mathbf{X}) := \operatorname{Cov}(\mathbf{X}, \mathbf{X}) = \begin{bmatrix}\operatorname{Var}(X_1) & \operatorname{Cov}(X_1, X_2) & \cdots & \operatorname{Cov}(X_1, X_k) \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Var}(X_2) & \cdots & \operatorname{Cov}(X_2, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_k, X_1) & \operatorname{Cov}(X_k, X_2) & \cdots & \operatorname{Var}(X_k)\end{bmatrix} = \mathbb{E}(\mathbf{X}\mathbf{X}^T) - \mathbb{E}(\mathbf{X})\left\{\mathbb{E}(\mathbf{X})\right\}^T.$$

Linearity of expectation holds for random vectors and matrices — we will not prove this here, but you may wish to tackle this as an exercise. Specifically, if $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{p\times q}$ are fixed matrices, and $\mathbb{M}_1$, $\mathbb{M}_2$ are random $(n \times p)$ matrices, then

$$\mathbb{E}(A\mathbb{M}_1 B) = A\mathbb{E}(\mathbb{M}_1)B, \text{ and } \mathbb{E}(\mathbb{M}_1 + \mathbb{M}_2) = \mathbb{E}(\mathbb{M}_1) + \mathbb{E}(\mathbb{M}_2).$$

One consequence of this is if $\mathbf{X}$ and $\mathbf{Y}$ are random vectors, $A$, $B$ conformable fixed matrices, and $c$, $d$ conformable fixed vectors, then

$$\text{Cov}(A\mathbf{X} + c, B\mathbf{Y} + d) = A\text{Cov}(\mathbf{X}, \mathbf{Y})B^T.$$

**Exercise 4.** Let $\mathbb{M}$ be a square, $(n \times n)$, random matrix. Show that $\text{tr}(\mathbb{E}(\mathbb{M})) = \mathbb{E}(\text{tr}(\mathbb{M}))$. As a hint, note that the $(i, j)$-th element of a matrix $M \in \mathbb{R}^{n\times n}$ is given by $e_i^T M e_j$, where $e_i$, $e_j$ are the $i$-th and $j$-th canonical basis vectors of $\mathbb{R}^n$, respectively.

Observe that

$$\text{tr}(\mathbb{E}(\mathbb{M})) = \sum_{i=1}^{n} e_i^T \mathbb{E}(\mathbb{M}) e_i = \mathbb{E}\left(\sum_{i=1}^{n} e_i^T \mathbb{M} e_i\right) = \mathbb{E}(\text{tr}(\mathbb{M})).$$

Alternatively, as Beau pointed out, we can ignore the hint and directly use the definition of the quantities involved. Specifically, the expected value of the trace of a matrix is the expected value of the sum of its diagonals, which by linearity is the sum of the expected values of the diagonals. But this is simply the trace of the expected value matrix as we have defined it!

When studying the properties of the least squares estimator under the linear model, we will have to work with the distribution of quadratic forms. When $\mathbf{X}$ is a random $k$-vector, $M \in \mathbb{R}^{k\times k}$ a symmetric matrix, $\mathbf{X}^T M\mathbf{X}$ is said to be a *quadratic form* in $\mathbf{X}$.

**Exercise 5.** For such a quadratic form, show that

$$\mathbb{E}(\mathbf{X}^T M\mathbf{X}) = \text{tr}(M\Sigma_{\mathbf{X}}) + \{\mathbb{E}(\mathbf{X})\}^T M\mathbb{E}(\mathbf{X}),$$

where $\Sigma_{\mathbf{X}} = \text{Cov}(\mathbf{X})$. Note that the variance of a quadratic form is in terms of higher moments of components of $\mathbf{X}$. As a result, it is much more complicated in the general case, and we will instead typically work directly with the distribution of quadratic forms, under assumptions.

The key is to abuse the fact that the trace of a scalar (or a $1 \times 1$ matrix) is the scalar itself, then use the cyclic property of trace. Observe that

$$\begin{aligned}
\mathbb{E}(\mathbf{X}^T M\mathbf{X}) &= \mathbb{E}(\text{tr}(\mathbf{X}^T M\mathbf{X})), \text{ by Exercise 4,} \\
&= \mathbb{E}(\text{tr}(M\mathbf{X}\mathbf{X}^T)), \text{ by cyclic property,} \\
&= \text{tr}(M\mathbb{E}(\mathbf{X}\mathbf{X}^T)), \text{ by Exercise 4 and linearity} \\
&= \text{tr}\left(M\left\{\Sigma_{\mathbf{X}} + \mathbb{E}(\mathbf{X})\{\mathbb{E}(\mathbf{X})\}^T\right\}\right), \\
&= \text{tr}(M\Sigma_{\mathbf{X}}) + \text{tr}\left(M\mathbb{E}(\mathbf{X})\{\mathbb{E}(\mathbf{X})\}^T\right) \\
&= \text{tr}(M\Sigma_{\mathbf{X}}) + \{\mathbb{E}(\mathbf{X})\}^T M\mathbb{E}(\mathbf{X}), \text{ by cyclic property,}
\end{aligned}$$

which we wanted to show.

## 4 The multivariate normal distribution

The random vector $\mathbf{X} : \Omega \to \mathbb{R}^k$ is said to have a *multivariate normal* or *Gaussian* distribution if $\mathbf{a}^T\mathbf{X}$ has a univariate normal distribution, for all $\mathbf{a} \in \mathbb{R}^k$. Letting $\boldsymbol{\mu} = \mathbb{E}(\mathbf{X})$ and $\Sigma = \operatorname{Cov}(\mathbf{X})$, we write $\mathbf{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$. Of course, we might want to write down the joint density of this distribution, but we will avoid this for now as it involves matrix determinant which we have not yet introduced.

**Exercise 6.** Show that the characteristic function of $\mathbf{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$ is

$$\varphi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\left(e^{i\mathbf{t}^T\mathbf{X}}\right) = \exp\left(i\mathbf{t}^T\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}\right).$$

Show that this implies $B\mathbf{X} \sim \mathcal{N}_q(B\boldsymbol{\mu}, B\Sigma B^T)$, for fixed $B \in \mathbb{R}^{q\times k}$.

Recall that if $W \sim \mathcal{N}(\mu, \sigma^2)$, then the characteristic function of $W$ is $\varphi_W(t) = \exp\left(i\mu t - \frac{t^2\sigma^2}{2}\right)$, for any $t \in \mathbb{R}$. By definition of multivariate normality, with $\mathbf{t} \in \mathbb{R}^k$ fixed, $W := \mathbf{t}^T\mathbf{X} \sim \mathcal{N}(\mathbf{t}^T\boldsymbol{\mu}, \mathbf{t}^T\Sigma\mathbf{t})$. It follows that

$$\varphi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}(e^{iW}) = \varphi_W(1) = \exp\left(i\mathbf{t}^T\boldsymbol{\mu} - \frac{\mathbf{t}^T\Sigma\mathbf{t}}{2}\right),$$

as claimed. Now for fixed $B \in \mathbb{R}^{q\times k}$, and $\mathbf{s} \in \mathbb{R}^q$,

$$\varphi_{B\mathbf{X}}(\mathbf{s}) = \mathbb{E}\left(e^{i\mathbf{s}^T B\mathbf{X}}\right) = \varphi_{\mathbf{X}}\left(B^T\mathbf{s}\right) = \exp\left(i\mathbf{s}^T B\boldsymbol{\mu} - \frac{\mathbf{s}^T B\Sigma B^T\mathbf{s}}{2}\right),$$

so $B\mathbf{X} \sim \mathcal{N}_q(B\boldsymbol{\mu}, B\Sigma B^T)$.

Partition the random vector $\mathbf{X} = [\mathbf{X}_1^T\ \mathbf{X}_2^T]^T$, where $\mathbf{X}_1$ is a $p$-vector and $\mathbf{X}_2$ is a $q$-vector. Let $\boldsymbol{\mu}_j = \mathbb{E}(\mathbf{X}_j)$, for $j = 1, 2$, and let

$$\Sigma_{ij} = \operatorname{Cov}(\mathbf{X}_i, \mathbf{X}_j),$$

for $i, j \in \{1, 2\}$. Two fundamental facts from probability theory are

- $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2 \iff \Sigma_{12} = \mathbf{0}_{p\times q}$,
- $\mathbf{X}_2 \,|\, \mathbf{X}_1 \sim \mathcal{N}_q\left(\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{X}_1 - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$.

Finally, we briefly introduce quadratic forms for Gaussian vectors. For $Z_1, \ldots, Z_k \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$,

$$\sum_{j=1}^k Z_j^2 \sim \chi_k^2.$$

More generally, if we have $(Z_1, \ldots, Z_k)$ independent with $Z_j \sim \mathcal{N}(\mu_j, 1)$, for $j = 1, \ldots, k$, we write $W = \sum_{j=1}^k Z_j^2 \sim \chi_k^2(\delta)$, where $\delta = \sum_{j=1}^k \mu_j^2$. We say that $W$ has a *non-central chi-squared distribution* with $k$ degrees of freedom and non-centrality parameter $\delta$. If $\mathbf{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \sigma^2 I_k)$, we will see that, for symmetric matrix $A \in \mathbb{R}^{k\times k}$, $\frac{\mathbf{X}^T A\mathbf{X}}{\sigma^2} \sim \chi_r^2(\delta)$, with $\delta = \frac{\boldsymbol{\mu}^T A\boldsymbol{\mu}}{\sigma^2}$, iff $A$ is idempotent with rank $r$. To understand this fully, we will need to derive the *spectral decomposition* of symmetric matrix $A$. As a teaser question, how would you go about simulating from a general $\mathbf{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$, without using `mvrnorm`?