A. Matrix Theory

A1. Vector Spaces
- Inner product
- Cauchy-Schwarz inequality
- Subspaces
- Orthogonal bases
- Orthogonal complement: \( L(x) = L(XX') \)
- Grammians, eigenvalues
- \( L^2 \) space
- Euclidean heuristic
- \( Z \in \mathbb{R} \Rightarrow L(\cos(Z)) = \mathbb{R} \)

A2. Projections
- Gram-Schmidt orthogonalization and Triangular representation
- Projections & facts
- General Metric projections \( I \) and \( \hat{I} \)
- Linear projection:
- 4 facts
- Gram-Schmidt
- Triangular Representation \( X = \hat{P} \hat{T} \hat{I} \)
- Grammians and projections, \( E_{xx} = \hat{E}_{xx} + \hat{E}_{xx} \), partial covering matrix
- Non-orthogonal projections

A3. Determinants, Volumes, and Jacobians
- Definition and properties of determinants
- Volume interpretation
- Cauchy-Binet Theorem
- Partitioned grammians
- Jacobians and their volume ratios
- Multivariate polar coordinates
- Integral Jacobians
- Chain rule, Calculation tricks, Hui's Lemma

A4. Spectral decompositions, Inverses, and Pseudo-inverses
- Eigenvector decomposition, Trace, Vector interpretation, higher-order traces, best linear predictors, Matrix square roots
- The Singular Value Decomposition
- Pseudo-inverses, 5 facts, Solving linear equations: shortest solutions
- Geometric interpretation, \( X = \Sigma \mu \Lambda \) and \( \hat{X} = \Sigma \hat{\mu} \hat{\Lambda} \)
- Student Sam example
- Partitioned inverses, Woodbury's theorem
- (\( \Sigma^{-1} \)) \( \Sigma^{-1} \) standard interpretation
- Partial correlation, Fisher information
A. Matrix Theory

A.1. Vector Spaces

Euclidean Space $\mathbb{R}^p : x = \left(\begin{array}{c} x_1 \\ \vdots \\ x_p \end{array}\right)$, $x \in \mathbb{R}^p$

- **Unit rules**: $x + y = \left(\begin{array}{c} x_1 + y_1 \\ \vdots \\ x_p + y_p \end{array}\right)$, $c x = \left(\begin{array}{c} c x_1 \\ \vdots \\ c x_p \end{array}\right)$

**Inner Product** $\langle x, y \rangle = x^t y = \sum_{i=1}^{p} x_i y_i$  
Length $\|x\| = \langle x, x \rangle^{\frac{1}{2}} = \left(\sum_{i=1}^{p} x_i^2\right)^{\frac{1}{2}}$

- **Cauchy-Schwartz**: $\langle x, y \rangle^2 \leq \langle x, x \rangle \langle y, y \rangle$ (or $|x \cdot y| \leq \|x\| \cdot \|y\|$)

**Angle** angle between $x$ and $y$

$$\cos(\theta_{xy}) = \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|}$$

(or $x \cdot y = \|x\| \cdot \|y\| \cos(\theta_{xy})$)

**More General Inner Products** $\langle x, y \rangle = x^t A y$, $A$ positive semi-definite

- Cauchy-Schwartz still valid
Linear Subspaces \[ V = (v_1, v_2, v_2) \in \mathbb{R}^p \]

\[ \mathcal{L}_{\text{co1}}(V) = \left\{ v : v = \sum_{i=1}^{q} c_i v_i \right\} \]

If \( v \) linearly independent then \( \mathcal{L}_{\text{co1}}(V) \) is a linear subspace of \( \mathbb{R}^p \).

Orthogonal Basis can always find an orthogonal set \( V = (v_1, \ldots, v_k) \).

\[ P' P = I_p, \text{ and so } v \]

\[ V = \sum_{i=2}^{p} C_{p \times 2} \begin{pmatrix} v_2 \\ v_3 \\ \vdots \\ v_p \end{pmatrix} \]

\[ \text{Leg } \ y_i = V_i / \|V_i\| \quad y_2 = \frac{1}{\|v_2\|} V_2 \quad \text{when } \frac{1}{\|v_2\|} = \frac{V_2 / \|v_2\|}{\|V_2 / \|v_2\|} \quad \text{etc.} \]

Orthogonal Representation \( v \) in \( \mathcal{L}_{\text{co1}}(V) \), \( v = Vc \) then

\[ v = FCc = \sum_{i=1}^{q} c_i \]

\[ p, q \times q \]
Orthocomplement $\mathcal{L}^\perp_{\mathcal{L}}(V)$ is the set of vectors orthogonal to $\mathcal{L}_{\mathcal{L}}(V)$.

$\mathcal{L}^\perp_{\mathcal{L}}(V) = \{ y : \langle y, v \rangle = 0 \text{ for all } v \in \mathcal{L}_{\mathcal{L}}(V) \}.$

Remark: $\langle y, y \rangle > 0$ for $y = 1, 2, \ldots, q$.

**Theorem 1 (Cauchy)** Given $X = \begin{pmatrix} x_1 & x_2 & \cdots & x_p \end{pmatrix}$, then

$\mathcal{L}(X) = \mathcal{L}(XX')$.

Proof: $A = XX' = \begin{pmatrix} a_1 & a_2 & \cdots & a_p \end{pmatrix}$ has columns $a_i$ that are linearly independent.

Combination of $x_1, x_2, \ldots, x_p$ so $\mathcal{L}(A) \subseteq \mathcal{L}(X)$.

Therefore:

$A = 0 \Rightarrow \mathcal{L}(A) \subseteq \mathcal{L}(X)$.

(Why?) However,

$v'A = 0 \Rightarrow v'XX'v = 0 \Rightarrow \|v'X\|^2 = 0 \Rightarrow v'X = 0$.

So $v \in \mathcal{L}(A) \Rightarrow v \in \mathcal{L}(X) \Rightarrow \mathcal{L}(A) \subseteq \mathcal{L}(X) \Rightarrow \mathcal{L}(A) = \mathcal{L}(X)$.
Ranks
\[ \text{rank}(X) = \dim \mathcal{L}(X) \]

(a) \quad \text{rank}(X) \leq \text{rank}(X') \quad ["\text{# indep rows} \leq \text{# indep columns}""]

(b) \quad \text{If } B \text{ and } C \text{ non-singular: } \text{rank}(BXC) = \text{rank}(X)

(c) \quad \text{rank}(X) = \text{rank}(XX') = \text{rank}(X'X) \quad [\text{Theorem 1}]

**Gramian:**
\[ X = (x_1, \ldots, x_n) = (v_1', \ldots, v_p') \quad \text{gives the Gramian} \]

\[ G = XX' = (v_i, v_j') \]

[more general:]
\[ C = XMX' = \langle v_i, v_j \rangle \quad \text{when } \langle v_i, v_j \rangle = v_i'Mv_j \]

- \( G \) is square, symmetric, positive semi-definite, \text{rank}(X)

- Eigen representation
\[ G = \mathbf{P} \mathbf{D} \mathbf{P}' = \sum_{i=1}^{\text{rank}(X)} \lambda_i x_i x_i' = \mathbf{H} \mathbf{H}' \quad \text{where} \]

\[ \mathbf{H} = (h_1, h_2, \ldots, h_p) \quad \text{with} \quad h_j = \nu_{j} v_j v_j' \]
H.Wrk 2/ H is a "square root of G" but it is not symmetric.

Find a symmetric square root.

$L^2$ Spaces  Suppose $Z(w) = \begin{pmatrix} Z_1(w) \\ \vdots \\ Z_p(w) \end{pmatrix}$, where $Z$ is a $p$-dimensional random variable on some $L^2$ space $L^2(\Omega, \mathcal{F}, P)$ with a prob distribution.

- For convenience assume expectation $\mathbb{E}[Z] = \mathbb{E}[Z(w)] dP(w) = 0$

- Covariance Matrix $\Sigma = \mathbb{E}[ZZ'] = \int Z(w)Z'(w) dP(w)$.

Linear Subspace $L(Z) = \{ X(w) = x'Z(w), x \in \mathbb{R}^p \}$, space of linear functions of $Z$. Notice $X = \Sigma x'$ has $\mathbb{E}[X] = 0$.

Inner Product on $L(Z)$: $X = x'Z$, $Y = y'Z$ then,

$\langle X, Y \rangle = \mathbb{Cov}(X, Y) = \mathbb{E}(X'XY)$

$= x' \Sigma y$. 
- Length \( \|X\|^2 = \langle X, X \rangle = x' \Sigma x = \text{Var}(X) \)

- Angle

\[
\cos(\theta) = \frac{\langle X, Y \rangle}{\|X\| \|Y\|} = \cos(\theta) \]

Homework 3: \( \mathbf{A} = \{a_1, a_2, \ldots, a_m \} \quad \mathbf{P} = (p_1, p_2, \ldots, p_m) \quad \sum p_i = 1 \)

\[
Z(w) = \frac{1}{n} \quad \text{with} \quad \sum Z(w) = 1
\]

That is, \( Z(w) \) is indicator vector of which of the \( n \) possible events has occurred. Give an explicit description of \( \langle X, Y \rangle \).

**Euclidean Heuristic**

Think of random variable \( X(w) \) as a vector \( \mathbf{X} = (X(w_1), X(w_2), \ldots, X(w_m)) \) in Euclidean space.
Similarly, given $X_1, \ldots, X_q \in \mathbb{R}(\Omega)$, think of $X_{1w} = \langle X_{1w}^2, \ldots, X_{qw}^2 \rangle$.

As a $q \times 2$ matrix,

$$A = \begin{pmatrix} X_1, & X_2, \\ \vdots, & \vdots, \\ X_q, & X_q \end{pmatrix}_{q \times 2}$$

Matrix multiplication is now defined using inner product $\langle \cdot, \cdot \rangle$,

so given

$$B = \begin{pmatrix} Y_1, & Y_2, \\ \vdots, & \vdots, \\ Y_q, & Y_q \end{pmatrix}_{q \times 2}$$

we get

$$AB' = \begin{pmatrix} \langle X_1, Y_1 \rangle, \ldots, \langle X_q, Y_q \rangle \end{pmatrix}_{q \times 1}$$

Main Point: All matrix results apply to probability calculations.

Hence $\sqrt{\mathbb{R}(\Omega)}$ is a linear subspace of $\mathbb{R}(\Omega, \mathbb{P})$. Argue that $\mathbb{R}(\Omega)$ is the set of random variables in $\mathbb{R}(\Omega, \mathbb{P})$ uncorrelated with $\mathbb{Z}$. 

\[ \mathbb{R}(\Omega) \]
Corollary to Theorem 1: Suppose \( Z(w) \) takes its values in a linear subspace \( \mathbb{R}^d \) of dimension \( q \leq p \). Then

\[
\text{Cov}(Z) = Z_{E_{	ext{cov}}}(Z) = Z_d.
\]

(Same result if \( Z(w) \) always in \( \mathbb{R}_+^d \).)

Factual Example: Multinomial Sampling,

a draws, \( L \) categories, \( \pi \) prob. \( \pi = (\pi_1, \pi_2, \ldots, \pi_L) \), \( y \sim \text{Mult}(n, \pi) \)

with \( y_k \) the number of draws in \( k \)th category. Then the vector of proportions \( p = (p_1, p_2, \ldots, p_L) \), with \( p_k = \frac{y_k}{n} \), takes its values in the simplex \( S_L \),

\[
S_L = \{v: v \geq 0, \frac{1}{L} \sum_k v_k = 1\}
\]

Hence \( S \) has covariance matrix \( \sum = \frac{1}{n} \text{(dist}(1-n)) \).

Why is this an example of the Corollary?
A.2 Projections, Gram–Schmidt Orthogonalization and Triangular Representation

Projections \( V = (v_1, v_2, \ldots, v_p) \) of rank \( p \):

\[
V = (v_1, v_2, \ldots, v_p)
\]

Wish to decompose \( y \) into

\[
y = \hat{y} + \hat{y}' = \sum_{\hat{y} \in \mathcal{L}_{2\infty}(V)} \frac{\hat{y}}{2} e^{\hat{y} \in \mathcal{L}_{2\infty}(V)}
\]

Write

\[
\hat{y} = V \hat{\beta}
\]

Then since

\[
V' y = V' (\hat{y} + \hat{y}') = V' \hat{y} = V' V \hat{\beta}
\]

we have

\[
\hat{\beta} = (V' V)^{-1} V' y
\]

(The matrix \( V' V \) is of rank \( \text{rank}(V) = p \) since it is a Grammian.)
Classic Example \( V = 1 \)

\[
\hat{P} = \frac{1}{n} \left( \begin{array}{c}
1' \\
1' \\
\vdots \\
1'
\end{array} \right) = \frac{1}{n} \hat{y}
\]

\[
\hat{P} = I - \frac{1}{n} \hat{y}
\]

For \( y = \left( y_1, y_2, \ldots, y_n \right)' \),

\[
\hat{y} = \hat{P} y = \frac{1}{n} \bar{y}
\]

\[
\bar{y} = y - \frac{1}{n} \bar{y} = (y_1 - \bar{y}, \ldots, y_n - \bar{y})'
\]

Two main statistical uses of projections:

(1) Get rid of noise (\( \bar{y} \) estimates true mean)

(2) Get rid of signal so we can examine the noise:

\[
\hat{\sigma}^2 = \frac{\sum y_i^2}{n}
\]
\[ \hat{y} = V(V'V)\hat{V}' y \quad \text{and} \quad \hat{y} = (I - V(V'V)^*V')y. \]

\[ \hat{P} \]

- \( \hat{P} \) is symmetric.
- \( \hat{P} \) is projection matrix into \( \mathbb{R}^p \), and \( \hat{P} \) is projection matrix into \( \mathbb{R}^p \).

Insert A.2.1

Some Facts About Projections

(i) \( \hat{P}^2 = \hat{P} \) ("Idempotent") \[ \text{since } \hat{P} \hat{y} = \hat{y} \]

(ii) \( \mathbb{R}^p(\hat{P}) = \mathbb{R}^p(V) \) \[ \mathbb{R}^p(\hat{P}) \subseteq \mathbb{R}^p(V) \] \[ \text{but } \hat{P}y = y, \quad y \in \mathbb{R}^p \]

\[ \Rightarrow \mathbb{R}^p(\hat{P}) \supseteq \mathbb{R}^p(V) \]

(iii) \( \text{rank}(\hat{P}) = p \quad \text{rank}(\hat{P}) = n - p \)

(iv) \( \hat{y}'\hat{y} = 0 \) \[ \hat{P}^2 = \hat{P}(I - \hat{P}) = \hat{P} - \hat{P}^2 = 0 \]
\( v \)  \( \| y \|_p^2 = y^T \hat{P}^T y = (y'V)(V'V)(V'y) \\
and \\
(\nu) \quad \| y \|_2^2 = \sum \| y \|_2^2 + \| y \|_2^2 \\
(\nu) \quad \| y \|_2 = \min_{v \in \mathbb{D}^P(V)} \| y - v \|_2^2 \\
(\nu i) \quad \cos^2 (a'\phi) = \frac{\| y \|_2^2}{\| y \|_2^2} = \frac{y^T \hat{P} \hat{P}}{y^T y} \\
\quad a'\phi = \min_{v \in \mathbb{D}^P(V)} a'\phi \\
(\nu i i) \quad \text{Suppose } V = \sum_{\phi \in \mathcal{P}} \sum_{\phi' \in \mathcal{P}'} \phi \phi' \phi' \phi' = I \\
\quad \text{Then} \\
\quad \hat{P} = \sum_{\phi \in \mathcal{P}} \sum_{\phi' \in \mathcal{P}'} \phi \phi' \phi' \phi' = PP' \\
\quad \hat{y} = \hat{P} y = PP'y \quad [\text{Note } \hat{P} \text{ axes align } y \text{ in } (x - \delta_p) \text{ coordinate system.}]
(iv) All eigenvalues of $A$ are either 0 or 1

\[ \hat{A}v = v \text{ for } v \in \mathcal{L}_{\text{col}}(V), \quad \hat{A}v = 0 \text{ for } v \notin \mathcal{L}_{\text{col}}(V) \]

\[ \hat{A} \text{ is identity matrix in } \mathcal{L}_{\text{col}}(V). \]

**Exercise 1**

Suppose $\mathcal{L}_{\text{col}}(V) = \mathcal{L}_{\text{col}}(\Gamma) \oplus \text{orthog} \mathcal{L}_{\text{col}}(\Gamma')$

When $A$ is orthogonal, $A^2 = I$. Show that $R$ is a

rotation matrix in $\mathcal{L}_{\text{col}}(V)$; that is $\hat{v} = Rv \in \mathcal{L}_{\text{col}}(V)$

and $\|\hat{v}\| = \|v\|$.

**Proceedings in General Metrics**

$\langle y, v \rangle = y^T A x$, $A > 0$. Then

$y = \hat{y} + \frac{1}{2} \hat{y}$
\[ \hat{y} = \left[ V \left( V^T A V \right)^{-} V^T \right] A y \]

\[ \hat{y} = \left[ A^T - V \left( V^T A V \right)^{-} V^T \right] A y = y - \hat{y} \]

\[ \begin{align*}
\text{Some Properties, e.g. } & \quad \hat{y}^T \hat{y} = y^T A \hat{y}, \\
& \quad \begin{align*}
\hat{y}^T \hat{y} &= \| \hat{y} \|^2 = y^T A y, \\
\| y \|^2 &= \| \hat{y} \|^2 + \| \hat{y}^T \|^2, \\
& \quad \Rightarrow \begin{align*}
\langle \hat{y}, \hat{y}^T \rangle &= 0, \\
\langle y, y^T \rangle &= \langle y, y \rangle - \langle y, y^T \rangle < y, y^T \rangle
\end{align*}
\end{align*}
\]

Homework 2: Verify \( \hat{y} \).

- \( \hat{y} = \hat{y}^T A y \) looks like one \( \hat{A} \) too many, but equivalently

\[ \hat{y} = V \langle V^T y \rangle \langle V^T \rangle \quad \text{like } \hat{y} = V \left( V^T V \right)^{-} V^T y \]

Homework 3: Show that the mapping \( \hat{V} = A^T V \), \( A^T \) symmetric square root of \( A \), gets us back to the case \( A = I \) for projection formulas.
Application to Linear Prediction  
Random Variables \( Y(w), V(w), \tilde{V}(w) \) in \( \mathbb{R}^3 \), all with expectation 0.

\[ \langle Y, V \rangle = \text{cov}(Y, V) = \text{E}Y\tilde{V} \]

Think of \( Y \) as \( Y = \left( Y(w) \right) \) and \( V \) as \( V = \left( V(w), V(w) - \tilde{V}(w) \right) \).

Projection Formulas

\[
Y = \frac{V \langle V, V \rangle \langle V, Y \rangle}{\langle V, V \rangle^2 \langle V, V \rangle} = \frac{\langle V, Y \rangle}{\langle V, V \rangle}
\]

when \( \Sigma = \begin{pmatrix} \sigma_{vv} & \sigma_{vy} \\ \sigma_{vy} & \sigma_{yy} \end{pmatrix} \) is joint covariance matrix of \( (Y, V) \).

More familiarly: \( \hat{y} = \sigma_{vy} \sigma_{vv}^{-1} \langle \hat{y}, v \rangle \).

Some Facts:
1. \( \hat{y} \) uncorrelated with \( \hat{y} = Y - \hat{Y} \) \[ \langle \hat{Y}, \hat{Y} \rangle = 0 \]
(iii) \( \text{Var} \hat{Y} = \hat{\sigma}_y^2 - \frac{\hat{\sigma}_v^2 \hat{\sigma}_{vy}}{\hat{\sigma}_v^2} \) 

\[ \| \hat{Y} \| = \| Y \| - \| \hat{Y} \| \]

\[ = \min_{6 \in \mathbb{R}^p} E \left( Y - \sum_{i=1}^{p} 6_i Y \right)^2 \]

(iv) \[ \frac{\text{Var} \hat{Y}}{\text{Var} Y} = \frac{\hat{\sigma}_v^2 \hat{\sigma}_{vy}}{\hat{\sigma}_y^2} = \hat{\beta} \frac{\sigma_Y}{\sigma_y} \]

\[ \text{multiple correlation coefficient between } Y \text{ and } V, \hat{Y}, V \]

Hence \( \hat{Y} / Y \) is the linear combination of \( V_i - V \), 

correlated with \( Y \). Show this and discuss the connection 

with property (iii) earlier.
Gram-Schmidt Orthogonalization \[ V = (v_1, v_2, \ldots, v_p) \] \[ \text{rank } p \]

Define
\[ V_j = (v_{i_1}, v_{i_2}, \ldots, v_{i_j}) \]
\[ P_j = V_j (V_j V_j)^{-1} V_j' \]

for \( j = 1, 2, \ldots, p \),

and let \( w_j = \frac{v_j}{\|v_j\|} \)

and
\[ \frac{1}{\sqrt{2}} V_2 = (I - P_1) V_2 \]
\[ w_2 = \frac{1}{\sqrt{2}} \frac{V_2}{\|V_2\|} \]
\[ \frac{1}{\sqrt{3}} V_3 = (I - P_2) V_3 \]
\[ w_3 = \frac{1}{\sqrt{3}} \frac{V_3}{\|V_3\|} \]

\[ \vdots \]

\[ \frac{1}{\sqrt{p}} V_p = (I - P_{p-1}) V_p \]
\[ w_p = \frac{1}{\sqrt{p}} \frac{V_p}{\|V_p\|} \]

Then \( W \) is orthonormal, \( W' W = I_p \), and

\[ \ell(v_i, v_j - w_j) = \ell(v_i, v_j') \quad j = 1, 2, \ldots, p. \]

From this we see that we can write
\[ v_1 = t_{11} w_1, \]
\[ v_2 = t_{22} w_1 + t_{22} w_2, \]
\[ v_3 = t_{13} w_1 + t_{23} w_2 + t_{33} w_3, \]
\[ \vdots \]
\[ v_p = t_{1p} w_1 + t_{2p} w_2 + \ldots + t_{pp} w_p, \]

when
\[ t_{kj} = w_k \cdot v_j, \]

In other words, \((t_{11}, \ldots, t_{pp})\) are the coordinates of \(v_j\) in the orthonormal basis \(w_1, w_2, \ldots, w_p\). In particular
\[ t_{jj} = w_j \cdot v_j = \frac{1}{\|v_j\|^2} v_j = \frac{1}{\|v_j\|} v_j \geq 0. \]

Thus
\[ \mathbf{V} = \mathbf{W} \mathbf{T} \]
\[ \text{orthogonal} \]
\[ \mathbf{V} = \mathbf{W} \mathbf{T} \]
\[ \text{upper triangle} \]
\[ \text{of the} \]
\[ \text{V} \]
Handout 5
(a) Show that $y$ is a linear function of $v_i - y$
and so $W = V^T$. (b) Show in general that $inv(A)$

an upper triangular matrix is upper triangular.

**Grammars and Projections**

\[ V = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \end{pmatrix} \quad V' = \begin{pmatrix} v_1' \\ v_2' \\ \vdots \end{pmatrix} \quad V = \begin{pmatrix} v_1 & v_2 \end{pmatrix} \]

\[ L = V'V = \begin{pmatrix} v_1' \\ v_2' \end{pmatrix} \begin{pmatrix} v_1 & v_2 \\ v_3 & v_4 \end{pmatrix} = \begin{pmatrix} v_1'v_1 & v_1'v_2 \\ v_2'v_1 & v_2'v_2 \end{pmatrix} \]

\[ = \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} \]

Now let \( \hat{p} = \frac{1}{\Lambda} \sum_i \hat{v}_i \) and \( \hat{p} = \frac{1}{\Lambda} \sum_i V_i \hat{p}_i \),

and \( \hat{V}_2 = \hat{p}_1 V_2 \quad \hat{V}_2 = \frac{1}{\hat{p}_1} V_2 \).
Since \( \mathbf{V}_2 \mathbf{V}'_2 = 0 \),

\[
\mathbf{L}_{22} = (\mathbf{V}'_2 \mathbf{V}_2) (\mathbf{V}'_2 \mathbf{V}_2) = \mathbf{V}'_2 \mathbf{V}_2 + \mathbf{V}'_2 \mathbf{V}_2
\]

Where

\[
\mathbf{V}'_2 \mathbf{V}_2 = \mathbf{V}'_2 \mathbf{P}'_2 \mathbf{V}_2 = \mathbf{V}'_2 \mathbf{V}_2 \mathbf{L}_{11}' \mathbf{V}'_2 \mathbf{V}_2
\]

\[
= \mathbf{L}_{21}' \mathbf{L}_{11}' \mathbf{L}_{12}
\]

Therefore,

\[
\mathbf{L}_{22}' = \mathbf{L}_{22} - \mathbf{L}_{21}' \mathbf{L}_{11}' \mathbf{L}_{12}
\]

Theorem A2/ (Pythagorean Extended) \( \mathbf{L}_{22} = \mathbf{L}_{22}' + \mathbf{L}_{22}'' \)

When

\[
\mathbf{L}_{22}' = \mathbf{L}_{22} - \mathbf{L}_{21}' \mathbf{L}_{11}' \mathbf{L}_{12} \text{ is Grammian of } \mathbf{V}_2 \text{ projected into } \mathbf{L}_{22}'' \mathbf{V}_2
\]

\[
\mathbf{L}_{22}'' = \mathbf{L}_{22} - \mathbf{L}_{21}' \mathbf{L}_{11}' \mathbf{L}_{12} \text{ is Grammian of } \mathbf{V}_2 \text{ orthogonal to } \mathbf{L}_{22}'' \mathbf{V}_2.
\]

Homework 6/ What does the Theorem say if \( p_1 = p_2 = 1 \)?
Hwk 7

\( \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \) has covariance matrix

\[
\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.
\]

Then the "partial covariance matrix of \( Y_2 \) after linear regression on \( Y_1 \)" is

\[
\frac{1}{\Sigma_{22}} \equiv \frac{1}{\Sigma_{22}} - \frac{1}{\Sigma_{21}} \Sigma_{11}^{-1} \Sigma_{12}.
\]

* Show that \( \frac{1}{\Sigma_{22}} \) is the covariance matrix of \( \hat{Y}_2 = Y_2 - \hat{\beta}_2 \hat{Y}_1 \), the residual of \( Y_2 \) after linear regression on \( Y_1 \).

* Interpret \( \frac{1}{\Sigma_{22}} \) in terms of Thm. A2.

Note \( \frac{1}{\Sigma_{22}} \) is the partial correlation coefficient of \( Y_1 \) with \( Y_2 \), partialing out \( (Y_1, Y_3, \ldots) \).
Non-Orthogonal Projections

(See Cadwell & Bell, Biometrics 48, 710-716)

- $A, B \in \mathbb{R}^{n \times k}$, rank $k$:

$$P = A (B^T A^T)^{-1} B^T$$

is "projection into $L(A)$ along $L(B)$";

- $(a)$ $P^2 = A (B^T A^T)^{-1} B^T A (B^T A^T)^{-1} B^T = P$ (Idempotent)
- $(b)$ $Pv = A (B^T A^T)^{-1} B^T v \in L(A)$
- $(c)$ $v \in L(B) \Rightarrow B^T v = 0$ and $\frac{1}{k} \sum_i v_i = 0$

- If $A = B$ then $P$ is orthogonal projection.

Homework: Suppose $k = 1$ and both $A$ and $B$ have unit length. Draw a picture illustrating how $P$ works.

Homework: Given $X$ of full rank, define $Q = X^T X^{-1}$ and

$$Q = I - \sum_{k=1}^n \frac{1}{k!} i! i! \left( \begin{array}{c} 1 \\ i \\ i \\ 1 \\ i \\ i \end{array} \right)$$

(a) Describe $Q$ as a $k$-dimensional projection matrix

(b) Show that $XQ$ has all rows with mean $0$
A3. Determinants, Volumes, and Jacobians

Traditional Definition \( A = (a_{ij}) \) has determinant

\[
|A| = \sum_{\pi} \text{sgn}(\pi) a_{\pi_1, \pi_2} \cdots a_{\pi_p} \quad [\pi \text{ permutation of } 1, 2, \ldots, p] 
\]

\[ p \leq 2 : |A| = a_{11}a_{22} - a_{12}a_{21} \]

- Original definition seldom used. Various algebraic properties substituted.

\[
(\text{i}) \quad \text{If } T \text{ is orthogonal, } |T| = \prod_{i=1}^{p} \lambda_i \quad (\Rightarrow |T| = 1) 
\]

\[
(\text{iii}) \quad |A| = |A'| 
\]

\[
(\text{viii}) \quad |a_1, a_2, \ldots, a_j, \ldots, a_p| < |A| 
\]

\[
(\text{vii}) \quad |a_1, a_2, \ldots, a_j + a_j', \ldots, a_p| = |A| \quad (j' \neq j) 
\]

\[
(\text{v}) \quad |AB| = |A||B| \quad [\Rightarrow |A'| = 1/|A|] 
\]

\[
(\text{vii}) \quad |A| = 0 \quad \text{iff } \text{rank}(A) < p \quad (1, 0 \leq \text{dim } R_0(A) < p) 
\]
\[ \begin{pmatrix} A & B \\ C & D \end{pmatrix} = AB \]

\[ e^{P^T} = I_p \]

\[ \text{Hankel: Use (v) to show that } \Gamma \text{ and } \text{ has } \Gamma = 1 \]

(a) If \( \Gamma \) is full rank, use \( \psi \) to prove (vii)

Volume Interpretation

Write \( A = W^T \)

Recall \( t_j = \| a_j \| \), the length of the

Attaching projection \( \psi \) out of \( \mathbb{R}^4 \)

Theorem

\[ |A| = |W| |T| = \pm \prod_j t_j \]

\[ = \prod_j \frac{\| a_j \|}{\| a_j \|} \]
Therefore

\[ \pm |A| = p \text{-dimensional volume of parallelepiped} \]

having its lower corner determined by \( a_1, a_2, \cdots, a_p \).

The sign of \( |A| \) depends on orientation of \( a_1, a_2, \cdots, a_p \) in picture.

Reversing roles of \( a_1, a_2 \) makes \( \frac{1}{2} \) go in wrong direction.

**Hence 2**/ For \( a_1 \) and \( a_2 \) in \( \mathbb{R}^2 \), give a geometrical interpretation of \( A(a_1, a_2) = \frac{1}{2} a_1 \cdot a_2 \).

**Hence 3**/ Let \( S \) be the simplex defined by the convex hull of \( 0 \) and \( a_1, a_2, \cdots, a_p \). Show that \( \text{vol}(S) = \pm \frac{1}{p!} a_1 \cdot a_2 \cdot \cdots \cdot a_p \).

(\( \text{Hint}: \) \( S = \{ v \in \mathbb{R}^p : v = \sum_{i=1}^{p} x_i a_i, \ x_i \geq 0 \text{ and } \sum_{i=1}^{p} x_i = 1 \} \))
\textbf{Cayley-Binet Theorem} \quad C = A \cdot B \\
\sum_{p \leq n} \frac{1}{p^2} - \frac{1}{p^2} \\
\text{Then} \quad \det C = \sum_{\text{subsets}} \prod_{i=1}^{p} \left| \begin{array}{ccc}
\frac{a_i}{\sqrt{p}} & \frac{a_i}{\sqrt{p}} & \cdots & \frac{a_i}{\sqrt{p}} \\
\frac{a_i}{\sqrt{p}} & \frac{a_i}{\sqrt{p}} & \cdots & \frac{a_i}{\sqrt{p}} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{a_i}{\sqrt{p}} & \frac{a_i}{\sqrt{p}} & \cdots & \frac{a_i}{\sqrt{p}}
\end{array} \right|

\text{Corollary} \quad C = A \cdot A' \quad \text{has} \\
\det C = \prod_{p \leq n} \left( 1 - \frac{1}{p^2} \right)

\textbf{Volume Interpretation} (Pythagoras extended)

\text{For } p=1: A = (a_1, a_2, \cdots, a_n) \text{ the corollary gives } \text{ length}\^2

\| A \|^2 = \sum_{i=1}^{n} a_i^2 = |C|
Hence 3.5] Suppose \( A = \begin{pmatrix} -v_1 & \vdots & -v_p \end{pmatrix} \).

Show that \( \| A \|_2 \) is determined by \( \langle v_i, v_j \rangle \) equaling the squared volume of the parallelepiped determined by \( v_i, v_j \).


\[ \text{Hint: use } A = W^T \text{ with } W^T = \frac{1}{\| v \|} \begin{pmatrix} v & \cdots & v \end{pmatrix}. \]

So the corollary says that the "squared volume of \( A \)" is the sum of all the projected squared volumes onto perpendicular coordinate axes.

\[
\begin{array}{cc|c}
A & B & \begin{cases} \| D \| \| A - BD^T C \| \text{ if rank}(D) = p \\
\| A \| \| D - C A B \| \text{ if rank}(A) = p \\
1 \end{cases} \\
& & 1
\end{array}
\]

Proof: \[
\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} A - BD^T C & B \\ 0 & C \end{pmatrix} = 1 \| D \| \| A - BD^T C \|.
\]
Partitioned Grammian \[ G = \begin{pmatrix} V_1 \end{pmatrix} \begin{pmatrix} V_1^T & V_2 \end{pmatrix} = \begin{pmatrix} f_1^{p_1 \times n_1} & f_1^{p_2 \times n_2} \end{pmatrix} \begin{pmatrix} f_2^{n_1 \times p_1} & f_2^{n_2 \times p_2} \end{pmatrix} \]

Applying Fact (viii)

\[ |G| = \left| \begin{array}{cc} f_1 & f_1' \\ f_2 & f_2' \end{array} \right| = \left| \begin{array}{cc} f_1 & f_1' \\ f_2 & f_2' \end{array} \right| \]

\[ = \left| V_1 V_2^T \right| = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \]

\[ \left( V_{11}, V_{12} \right)^2 = \left( V_{11}, V_{12} \right)^2 \left( V_{11}, V_{12} \right)^2 \]

\[ \text{Hence \ 4/ \ Draw \ a \ picture \ of \ the \ case \ } p_1 = 2 \quad p_2 = 1. \]

**Fact (ix)** \[ |I_p + AB| = |I_p + BA| \]

**Proof** \[ \begin{vmatrix} I_p - A \end{vmatrix} = \begin{vmatrix} I_p - A \end{vmatrix} = |I_p + A^T B| \]

**Hence 5/ \ Find** \[ \begin{vmatrix} I_p - A \end{vmatrix} = \begin{vmatrix} 0 \end{vmatrix} \]
Jacobian

Suppose \( \hat{X} = m(x) \) maps \( \mathbb{R}^n \) \( \leftrightarrow \) \( \mathbb{R}^2 \)

in a continuously differentiable, invertible way, and

that \( f(x) \) is the density of \( x \). What is the density

of \( \hat{x} \), say \( \hat{f}(\hat{x}) \)?

Answer

Let \( M(x) = \left( \begin{array}{c} \frac{\partial \hat{x}_1}{\partial x_1} \\ \vdots \\ \frac{\partial \hat{x}_n}{\partial x_n} \end{array} \right) = \frac{d\hat{x}}{dx} \), so \( M(x) = \left( \begin{array}{c} \frac{\partial \hat{x}_1}{\partial x_1} \\ \vdots \\ \frac{\partial \hat{x}_n}{\partial x_n} \end{array} \right) = \frac{d\hat{x}}{dx} \).

Thm

\[ f(\hat{x}) = f(x)|M(x)| \]

\[ = f(x)|J(x \rightarrow \hat{x})| \]

\[ \frac{|Jacobi|}{|Jacobi|} \]

\[ \frac{dx}{d\hat{x}} \]

Why?

Infinitesimal Volume \( dA \) becomes

\[ x + dx = x + \sum_{i=1}^{n} c_i dy_i \]
maps locally into

\[ \tilde{x} + M dx = \tilde{x} + (M_x M - M_x) dx = \tilde{x} + \Sigma M_i dx_i. \]

Parallelepiped \( d\tilde{A} \) has volume \( \|M\| dA \) compared to

volume \( \pi dx \) for \( dA \). Therefore the "conservation of probability" gives

\[ f(\tilde{x}) \|M\| dx \tilde{x} = f(x) \pi dx. \]

or

\[ f(\tilde{x}) = f(x) \|M\| J(x \to \tilde{x}) \]

\[ = P(x) J(x \to \tilde{x}) = P(x) \frac{dx}{d\tilde{x}} \] Easy to remember.

Note \( dA \) doesn't have to be a cubic.

From the other direction \[ J(\tilde{x} \to x) = \frac{1}{\sqrt{J(x \to \tilde{x})}} = \frac{1}{\left| \frac{dx}{d\tilde{x}} \right|} = \frac{d\tilde{x}}{dx} \]
Example: Multivariate Polar Coordinates

\[ x_1 = g \cos \theta_1 \]
\[ x_2 = g \sin \theta_1 \cos \theta_2 \]
\[ x_3 = g \sin \theta_1 \sin \theta_2 \cos \theta_3 \]
\[ \vdots \]
\[ x_n = g \sin \theta_1 \sin \theta_2 \cdots \sin \theta_{n-2} \cos \theta_{n-1} \]
\[ x_n = g \sin \theta_1 \sin \theta_2 \cdots \sin \theta_{n-2} \sin \theta_{n-1} \]

For

\[ 0 \leq g \]
\[ 0 \leq \theta_n < \pi \]
\[ \theta_k, \forall k = 1, 2, \ldots, n-1 \]
\[ 0 \leq \theta_i \leq 2\pi \]

\[ \text{Hnwk 6:} \]
(a) Show \[ J(x \rightarrow g, \theta_1 \ldots \theta_n) = g^n \left( \sin \theta_1 \right)^{n-2} \left( \sin \theta_2 \right) \cdots \left( \sin \theta_{n-1} \right) \]
(b) Show that the unit sphere in \( \mathbb{R}^n \) has (\( -1 \)) dimensional "area"

\[ A = 2\pi^{n/2} / \Gamma(n/2) \]

\[ \left( \text{Use Fact: } \int_0^{\pi/2} \sin^m(x) \, dx = \frac{\Gamma(m+1)}{2} \Gamma\left( \frac{m+1}{2} \right) \right) \]

For completeness' sake, unit sphere's volume = \( \frac{2\pi^{n/2}}{\Gamma(n/2)} = A / n \).
Integral Jacobians

Suppose \( x \in \mathbb{R}^n \) maps into \( y = (y_1, y_2) \in \mathbb{R}^2 \),
\[ y = (y_1, y_2) = \phi(x), \]
and that the density of \( x \) depends only on \( y \),
\[ f_X(x) = g(y). \]

What is the density of \( y_1 \)?

Answer
\[ f_{Y_1}(y_1) = \int_{y_2} g(y_1, y_2) \, dx = g(y_1) \int_{y_2} J(x \rightarrow y_1, y_2) \, dy_2 \]

Definition
The integral Jacobian of \( x \rightarrow y \) is
\[ J(x \rightarrow y) = \int_{y_2} J(x \rightarrow y_1, y_2) \, dy_2. \]

Lemma
If \( f_X(x) = g(y) \) then \( f_{Y_1}(y) = g(y) \cdot J(x \rightarrow y) \).
Lemma 2 \[ J(x \rightarrow y) = \lim_{\Delta x \to 0} \frac{Vol_{\Delta x}^3(x, y, x_1 + \Delta x)}{Vol_{\Delta x}^3(y, y_1, \Delta x)} \]

Thus \( J(x \rightarrow y) \) does not depend on \( f(x) \) or the choice of \( y_1 \) to "complete" \( y \).

**Homework 7//** Give a "conservation of probability" argument for Lemma 2.

Lemma 3 \[ J(x \rightarrow y_1, y_2) \text{ and } y \rightarrow (y_1, y_2) \text{ then the chain rule holds,} \]

\[ J(x \rightarrow y_1, y_2) = J(x \rightarrow y_1) J(y_1 \rightarrow y_2) \]

**Homework 8//** Give a "local volumes" argument for Lemma 3.
Example: Polar Coordinates

\[ \mathbf{L}(x \to \theta) = \int_{\theta_1}^{\theta_2} \mathbf{P}(\mathbf{L}^{-1}(\theta)) \, d\theta \]

\[ = \int_{\theta_1}^{\theta_2} \mathbf{P} \, d\theta \]

\[ = \int_{\theta_1}^{\theta_2} \mathbf{P} \, d\theta \]

\[ = 2\pi \mathbf{P} \]

\[ \mathbf{L}(x \to \theta) = 2\pi \mathbf{P} \]

\[ \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \]

\[ f(x) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}||x||^2} = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}S^2} \]

\[ f(S) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}S^2} \frac{1}{\Gamma(n/2)} \]

\[ f(S) = \frac{1}{\Gamma(n/2)} e^{-\frac{1}{2}S^2} \]

[ Corollary: \( S \sim \sqrt{\frac{X^2}{\mathbf{P}}} \)]
Trick for Calculating $J(x \to y)$: Reverse Lemma 1.

Since $J(x \to y) = \frac{f_Y(y)}{g_x(x)}$, we can calculate $J(x \to y)$ without integration if we can find any such distribution where we know $f_Y(y) / f_x(x)$.

Example: $\Omega_x = \{ x : x \geq 0 \}$, $x = 1, 2, \ldots, 3$ and $S = \sum x_i$, $\Omega_S = [0, \infty)$

What is $J(x \to S)$?

Consider $X_i \sim \mathcal{G}_x$, so $f_X(x) = \prod e^{-x_i} = e^{-S}$.

Then $S = \sum X_i \sim \mathcal{G}_x$, with $f_S(x) = \frac{x^{n-1} e^{-x/a}}{a^n}$. Therefore

$$J(x \to S) = \frac{x^{n-1} e^{-x/a}}{a^n} = \frac{x^n}{e^{x/a}}$$
Homework 9/ Use this result to prove that the n-dimensional volume of \( B_1 = \frac{\sqrt{n}}{\sqrt{\pi}} \), where \( \sum x_i = 1 \).

Homework 10/ \( \mathbb{R}^n = \mathbb{R}^n \), \( \sum x_i = 1 \). What is \( J(x \to y) \)?

**Hausdorff Lemma (Important Later):** \( x \to \hat{s} = \frac{x}{\|x\|} \)

\[
J(x \to \hat{s}) = \left( \frac{\|x\|}{\sqrt{n}} \right)^n \frac{\pi^{n-1}}{\Gamma(n/2)}
\]

Homework 11/ How does this map relate to the result

\[
J(x \to y) = 2\pi \frac{n^{-1}}{\Gamma(n/2)}
\] on 3.11?
Spectral Decomposition

A real, symmetric, rank(A) = n:

$$A = \Gamma \Lambda \Gamma^\top = \sum_{i=1}^{n} \lambda_i \delta_i \delta_i^\top \quad (\lambda \neq 0)$$

Where

$$\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)$$

is orthogonal, $\Gamma^\top \Gamma = I$, and

$\Lambda$ is diagonal, $\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \\ \vdots & \vdots \\ 0 & \lambda_n \end{pmatrix}$. Since

$$A \Gamma = \Gamma \Lambda \Leftrightarrow A \delta_i = \lambda_i \delta_i \quad (\lambda \neq 2 \cdot n)$$

we see that $\lambda$ are the eigenvalues ("characteristic values") of $A$ and $\delta_i$ the corresponding eigenvectors.

**Exercise 1** Show that $\lambda_{col}(A) = \lambda_{col}(\Gamma)$. 

\[1A41] \]
Matrix Trace

If \( A \) is \( p \times p \),

\[
\text{tr}(A) = \frac{p}{\lambda_i} a_{ii}.
\]

Fundamental Property

\[
\text{tr}(XY) = \text{tr}(YX)
\]

Vector Interpretation

Let \( Z = Y' \). Then

\[
\text{tr}(XY) = \text{tr}(XZ') = \sum_{i=1}^{p} \sum_{j=1}^{q} x_{ij} z_{ji} = \sum_{i=1}^{p} \sum_{j=1}^{q} x_{ij} z_{ji}
\]

where \( x \) and \( z \) are vectors

\[
= x \cdot \frac{z}{p}
\]

For Fundamental Property is just \( x \cdot \frac{z}{p} = z \cdot x \).

If \( A \) is symmetric

\[
\text{tr}(A) = \text{tr}(A^T A) = \text{tr}(A A') = \text{tr}(A^T) = \text{tr}(A).
\]

\[
= \sum_{i=1}^{p} \lambda_i.
\]
The trace of any square matrix $A$, $\text{tr}(A) = \sum_i a_{ii}$, holds for any $p \times p$ matrix $A$, though $\lambda_i$ may be complex. More general "traces" are possible: for $\tilde{s} = (i_1, i_2, \ldots, i_p)$, e.g., let $A_{\tilde{s}}$ be the submatrix of $A$ formed from just those rows and columns.

e.g. $A_{(1, 2)} = (a_{11}, a_{12}, a_{21}, a_{22})$. Then

$$\sum_{\tilde{s}} |A_{\tilde{s}}| = \sum_{\tilde{s}} \prod_{i=1}^p \lambda_{i}$$

Both sums over all subsets of $\{1, 2, \ldots, p\}$ of size $p$.

"Trace of $A$" is a fundamental symmetric function of eigenvalues.

Hence $\frac{1}{p} \sum_{i=1}^p a_{ii}$ symmetric represents a projection matrix into a $k$-dimensional subspace $\mathcal{L}$ of $\mathbb{R}^p$. Show that $\text{tr}(P) = k$. 

$\text{tr}(P) = k$. 
Problem 3/4 \(X\) and \(Y\) are random vectors with means \(\mu_x\) and \(\mu_y\) and covariance

\[
\begin{pmatrix}
X \\
Y
\end{pmatrix} \sim \begin{pmatrix}
\mu_X \\
\mu_Y
\end{pmatrix} \begin{pmatrix}
\Sigma_{xx} & \Sigma_{xy} \\
\Sigma_{yx} & \Sigma_{yy}
\end{pmatrix}
\]

Show that \(\hat{Y} = \mu_y + \Sigma_{yx} \Sigma_{xx}^{-1}(X - \mu_x)\) is the best linear predictor of \(Y\) in terms of \(X\), in the sense that it minimizes the trace of the covariance matrix of \(\hat{Y} = Y - \hat{Y}\) among all choices of \(M\).

Matrix Square Roots \(A\) symmetric, \(\geq 0\) (i.e., \(\lambda \geq 0\))
We can use $A = RL^2R'$ to define the matrix square root.

$$
S = RL^2R' \quad (L^2 = \begin{bmatrix} \sqrt{a} & 0 \\ 0 & -\sqrt{a} \end{bmatrix})
$$

So,

$$
S^2 = RL^2RRL^2R' = RL^4R' = A.
$$

Other square roots can be useful, for example

$$
\tilde{S} = \begin{bmatrix} L^2 & 0 \\ 0 & M^2 \end{bmatrix}
$$

satisfying

$$\tilde{S}^{-1}\tilde{S} = RL^2R' = A. \quad \text{We can pull out } \tilde{S} \text{ for}
$$

$$
\tilde{S} = \begin{bmatrix} \tilde{S} \\ \text{zeros} \end{bmatrix}
$$

still giving $\tilde{S}^{-1}\tilde{S} = A$. Then writing

$$
\tilde{S} = W T
$$

$$
A = \tilde{T}' W W T = \tilde{T}' T, \quad \tilde{T} \text{ is a triangular square root.}
$$
The Singular Value Decomposition: Any matrix \( X \) with \( \text{rank}(X) = m \) can be written as

\[
X = U \Sigma V^T = \sum_{i=1}^{\min(m,n)} \sigma_i u_i v_i^T
\]

where \( U^T U = V V^T = I \), and \( \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 \).

The \( \sigma_i \) are the singular values, columns of \( U \) are the left singular vectors, and columns of \( V \) are the right singular vectors.

Then \( X' X = V \Sigma U^T \Sigma U V^T = V \Sigma^2 V^T \), so \( A = X' X \) has

- eigenvalues \( \sigma_i^2 \) and eigenvectors columns of \( V \). Likewise,

\[
A = X' X = U \Sigma \Sigma U^T \text{ has } \text{eigenvalues } \sigma_i^2 \text{ and eigenvectors columns of } U.
\]

Hint 3.5: From \( \text{row basis} \) one gets coordinates of \( X \) row in the orthogonal basis \( V = (v_1, v_2, \ldots, v_n) \).

Likewise, \( \text{col basis} \) gives coordinates of \( X \) columns in orthonormal basis \( U = (u_1, u_2, \ldots, u_m) \).
Proof of SVD

\[ A = X'X = VD'V \quad (\text{so } d_i^2 = \lambda_i \quad V_i^2 = \lambda_i) \]

and

\[ U = XDV' \]

Then

\[ U'U = V'X'XV = V'D'V = D' = 0 \quad 0 \quad \ldots \quad I \]

and

\[ UDV' = XVVD'V' = XV' \]

Proof

\[ \rho_{col}(V) = \rho_{col}(A) = \rho_{row}(X') = \rho_{row}(X) \]

Easy

Theorem 3.3

By

\[ P = VV' \quad \text{is projection matrix with } \rho_{row}(X) \]

implies

\[ x^TP = x \quad \text{for each row of } X \Rightarrow XV'V = X \]
Mehler's Identity ("tetrahorn Series")

\[ f_{\nu}(x, y) = \sum_{j=0}^{\infty} \frac{(\nu(x))!}{j!} \frac{(\nu(y))!}{j!} \psi(j, x) \psi(j, y) \psi(j) \]

where \( \psi(j) \) is the \( j \)-th Hermite polynomial \( \psi(j)(x) = \frac{d^j e^{-x^2}}{dx^j} \)

(\( \nu = 3, \psi = \psi \)).

If \( x, y \) vectors, the Matrix

\[ f_{\nu}(x, y) = \sum_{j=0}^{\infty} \frac{(\nu(x))!}{j!} \frac{(\nu(y))!}{j!} \psi(j, x) \psi(j, y) \]

is outer product

and the right side is the sum of the Matrix.

Pearson, Fisher, Hobson, Correspondence And you -

H. C. Lancaster 1957 Biomath 239-242

Hand/Verify & numerically
Homework 4

Show (i) $L_{0,1}(X) = L_{0,1}(U)$ (ii) $L_{row}(X) = L_{0,1}(V)$

Pseudo-Inverses

If $A$ is of full rank, $A = P A P'$ gives

$$A' = P A P'$$

as its inverse. We can use the SVD $X = U D V'$ to define something that acts like an inverse for $X$.

(a) Pseudo-Inverse of $X$

$$X' = V D^{-1} U$$

Some Facts

(i) $L_{0,1}(X') = L_{0,1}(V) = L_{row}(X)$

(ii) $L_{row}(X') = L_{0,1}(U) = L_{0,1}(X)$
(iii) \( XX' = U D V' V D' U' = U U' = \mathbf{I}_p \)

= Projection Matrix into \( \mathbb{L}_p(X) \) (Identity matrix in \( \mathbb{L}_p(X) \))

(iv) \( X X' = V D U' U D V = V V' = \mathbf{I}_m \)

= Projection Matrix into \( \mathbb{L}_m(X) \) (Identity matrix in \( \mathbb{L}_m(X) \))

(v) \( X X' X = U D V' (V D U') V D V = U D V' \)

= \( X \) (Sometimes definition of Pseudo Invert.)

Solving Linear Equations With \( b \) solve for \( \beta \)

Given

\[ n = X \beta, \quad \beta \in \mathbb{L}_p(X), \]

Then

\[ \beta_* = X^{-} \in \mathbb{L}_p(X) \in \mathbb{L}_p(X), \text{ and} \]

is a solution since
\[ X_\beta = X X^T \gamma = P_{\text{row}} \gamma = \gamma. \]

Notice that there can only be one solution in \( \mathbb{R}^m \): if \( X \beta_i = \gamma \) also, then
\[ X (\beta - \beta_i) = \gamma - \gamma = 0 \quad \text{so} \quad \beta_i, \beta \in \mathcal{L}_{\text{row}}. \]

(vii) The general solution to \( \gamma = X \beta \) for \( \gamma \in \mathcal{L}_{\text{col}}(X) \) is \( \beta = X^{-1} \gamma \). Therefore
\[ \beta_0 = X^{-1} \gamma. \]

\( \beta_0 \) is the shortest solution to \( \gamma = X \beta \).

If we try to solve \( y = X \beta \) for \( y \) possibly
not in \( \mathcal{L}(X) \), then \( \hat{\beta} = X^+y \) given

\[
X\hat{\beta} = X^+y = P_{\mathcal{L}(X)} y = \hat{y},
\]

the projection of \( y \) \( \in \mathcal{L}(X) \) is

\[\hat{\beta}\] is the 

**Least Squares** solution for \( \beta \).

The **Full Rank Case** \( X \), \( p \leq n \), \( \text{rank}(X) = p \).

\[
X = UDV', \quad X = VDV', \quad [V \text{ orthonormal } p \times p]
\]

\[\text{row} \mathcal{L}(X) = \mathcal{L}_p(X) \text{ and}
\]

\[
X'X = VD'U'UDV' = VV' = I_p
\]

since \( VV' = V'V \) \( p \times p \) \text{ orthonormal}. 

```
If we write $X = \left( \frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_p} \right)$ and $X = \begin{pmatrix} -x_1^2 \\ \vdots \\ -x_p^2 \end{pmatrix}$, then this says

$$\frac{1}{x_1} \cdot x_1 = \frac{1}{x_2} \cdot x_2 = \cdots = \frac{1}{x_p} \cdot x_p = 1$$

**Picture for $p = 2$**

Let

$$\frac{1}{x_1} = x_1 - \frac{x_1 x_2}{11x_2^2} x_2 \quad \text{(part 1)} \quad \frac{1}{x_2} = x_2 - \frac{x_2 x_1}{11x_1^2} x_1 \quad \text{(part 2)}$$

Then $x_1^2 x_2 = 0$, so

$$x_1 = \frac{1}{x_1}$$

and

$$x_2 = \frac{1}{x_2}$$

and likewise.

**Hence**

In general

$$x_i = \frac{x_i}{11x_i^2}$$

where $\frac{x_i}{y_i}$ is the part of $x_i$ not in $x_i x_2, x_3, \ldots, x_p$. 
Notice that each $x_j$ is like the last vector in the Gram-Schmidt orthogonalization procedure.

Orthogonal Representation of $X$

$$X = UDV' = \sum_{k=1}^{p} u_k v_k$$

where $U = (u_1, \ldots, u_p)$ and $V = (v_1, \ldots, v_p)$. Now think

If $u_k v_k$ is a vector in $\mathbb{R}^p$ with coordinates

$$u_k v_k \in \mathbb{R}^p$$

Homework 6: Show that the $u_k$ are orthogonal vectors of unit length in $\mathbb{R}^p$. 

$$\left(\begin{array}{c}
\frac{1}{\|x_j\|} \\
\frac{x_j}{\|x_j\|^2}
\end{array}\right) = \text{Useful vector}$$
Similarly, thinking of $X$ as a vector $x \in \mathbb{R}^p$, 

$$x = \sum_{k \sim \lambda} \lambda_k \mathbf{u}_k \mathbf{v}_k^T$$

is the SVD representation. SVD programs (like "svd" in R) return $d_1 > d_2 \cdots > d_k > 0$. Often the first one or two or three terms are enough to give a good approximation to $X$, say 

$$\hat{X}(k) = \sum_{k=1}^K d_k \mathbf{u}_k \mathbf{v}_k^T, \quad k \leq K.$$ 

**Hence, Show that**

$$\|x - \hat{X}(k)\|^2 = \sum_{k=k}^\infty d_k^2$$

(so \(R^2 = \frac{\sum d_k^2}{\sum d_k^2}\))

Here is an example from Murdoch, Kent, and Bibby.
STUDENT SCORE DATA MARDIA, KENT, AND BIBBY

* 88 students each took 5 tests, either Close or Open book:

<table>
<thead>
<tr>
<th>student</th>
<th>mechC</th>
<th>vecC</th>
<th>algO</th>
<th>analyO</th>
<th>statO</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>77</td>
<td>82</td>
<td>67</td>
<td>67</td>
<td>81</td>
<td>374</td>
</tr>
<tr>
<td>2</td>
<td>63</td>
<td>78</td>
<td>80</td>
<td>70</td>
<td>81</td>
<td>372</td>
</tr>
<tr>
<td>3</td>
<td>75</td>
<td>73</td>
<td>71</td>
<td>66</td>
<td>81</td>
<td>366</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
<td>72</td>
<td>63</td>
<td>70</td>
<td>68</td>
<td>328</td>
</tr>
<tr>
<td>5</td>
<td>63</td>
<td>63</td>
<td>65</td>
<td>70</td>
<td>63</td>
<td>324</td>
</tr>
<tr>
<td>6</td>
<td>53</td>
<td>61</td>
<td>72</td>
<td>64</td>
<td>73</td>
<td>323</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>15</td>
<td>40</td>
<td>43</td>
<td>23</td>
<td>18</td>
<td>139</td>
</tr>
<tr>
<td>84</td>
<td>15</td>
<td>38</td>
<td>39</td>
<td>28</td>
<td>17</td>
<td>137</td>
</tr>
<tr>
<td>85</td>
<td>5</td>
<td>30</td>
<td>44</td>
<td>36</td>
<td>18</td>
<td>133</td>
</tr>
<tr>
<td>86</td>
<td>12</td>
<td>30</td>
<td>32</td>
<td>35</td>
<td>21</td>
<td>130</td>
</tr>
<tr>
<td>87</td>
<td>5</td>
<td>26</td>
<td>15</td>
<td>20</td>
<td>20</td>
<td>86</td>
</tr>
<tr>
<td>88</td>
<td>0</td>
<td>40</td>
<td>21</td>
<td>9</td>
<td>14</td>
<td>84</td>
</tr>
</tbody>
</table>

* I standardized the columns (to have mean 0, sd=1):

<table>
<thead>
<tr>
<th>mechC</th>
<th>vecC</th>
<th>algO</th>
<th>analyO</th>
<th>statO</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.18</td>
<td>2.39</td>
<td>1.54</td>
<td>1.37</td>
<td>2.24</td>
</tr>
<tr>
<td>1.38</td>
<td>2.08</td>
<td>2.77</td>
<td>1.57</td>
<td>2.24</td>
</tr>
<tr>
<td>2.06</td>
<td>1.70</td>
<td>1.92</td>
<td>1.30</td>
<td>2.24</td>
</tr>
<tr>
<td>0.92</td>
<td>1.63</td>
<td>1.17</td>
<td>1.57</td>
<td>1.49</td>
</tr>
<tr>
<td>1.38</td>
<td>0.94</td>
<td>1.36</td>
<td>1.57</td>
<td>1.20</td>
</tr>
<tr>
<td>0.80</td>
<td>0.79</td>
<td>2.01</td>
<td>1.17</td>
<td>1.78</td>
</tr>
<tr>
<td>-1.37</td>
<td>-0.81</td>
<td>-0.72</td>
<td>-1.60</td>
<td>-1.41</td>
</tr>
<tr>
<td>-1.37</td>
<td>-0.96</td>
<td>-1.09</td>
<td>-1.26</td>
<td>-1.47</td>
</tr>
<tr>
<td>-1.94</td>
<td>-1.57</td>
<td>-0.62</td>
<td>-0.72</td>
<td>-1.41</td>
</tr>
<tr>
<td>-1.54</td>
<td>-1.57</td>
<td>-1.75</td>
<td>-0.79</td>
<td>-1.23</td>
</tr>
<tr>
<td>-1.94</td>
<td>-1.87</td>
<td>-3.35</td>
<td>-1.80</td>
<td>-1.29</td>
</tr>
<tr>
<td>-2.23</td>
<td>-0.81</td>
<td>-2.79</td>
<td>-2.54</td>
<td>-1.64</td>
</tr>
</tbody>
</table>

* The singular value decomposition of the standardized matrix had

$$d = 16.64 \quad 8.02 \quad 6.22 \quad 5.80 \quad 4.63$$

and

$$u = \begin{bmatrix}
\{1\} & \{2\} & \{3\} & \{4\} & \{5\} \\
\{1\} & 0.400 & 0.646 & -0.621 & 0.141 & 0.131 \\
\{2\} & 0.431 & 0.440 & 0.707 & -0.294 & 0.182 \\
\{3\} & 0.503 & -0.129 & 0.037 & 0.110 & -0.847 \\
\{4\} & 0.457 & -0.388 & 0.131 & 0.667 & 0.423 \\
\{5\} & 0.438 & -0.470 & -0.309 & -0.651 & 0.233
\end{bmatrix}$$

with the first two columns of $$u$$:

1 0.258 0.084
2 0.273 -0.026
3 0.246 0.034
4 0.182 -0.019
5 0.173 -0.005
6 0.180 -0.085
83 -0.157 0.017
84 -0.164 0.002
85 -0.163 -0.115
86 -0.185 -0.072
87 -0.280 -0.042
88 -0.272 0.040
First two principal components $v_{[1]}, v_{[2]}$

First two loadings $u_{[1]}, u_{[2]}$
for each of the 88 students
(a) Use the R function "svd" to compute $\tilde{x}(1)$ and $\tilde{x}(2)$ for the "scored" data; plot $\tilde{x}(1)$ versus $x$ in a scattergram, and similarly $\tilde{x}(2)$ versus $x$.

(b) Compute $r_x \sim \tilde{x}(k)\tilde{x}(k)^T$ for $k=1,2$, and verify the relationship $r = \frac{\sum d_k^2}{\sum d_k^2}$.

Partitioned Inverse

For $p \times k$ non-singular

\[
G = \begin{pmatrix}
    G_{11} & G_{12} \\
    G_{21} & G_{22}
\end{pmatrix}
\]

\[H = G' = \begin{pmatrix}
    H_{11} & H_{12} \\
    H_{21} & H_{22}
\end{pmatrix},
\]

Then

\[
H = \begin{pmatrix}
    (G_{11} - G_{12} G_{22}^{-1} G_{21})' & G_{11}^{-1} (G_{22} - G_{21} G_{11}^{-1} G_{12})' \\
    -G_{22}^{-1} G_{21} (G_{11}^{-1} G_{12})' & (G_{22} - G_{21} G_{11}^{-1} G_{12})'
\end{pmatrix}
\]
proved directly by verifying $\mathbf{LH} = \mathbf{I}$. Special case of

**Woodbury's Theorem**: $A$ and $B$ non-singular, then

\[
(A + \mu B^t B)^{-1} = A^{-1} - A^{-1} B (V^t B^t + \mu V^t A^{-1} V)^{-1} V^t A^{-1}
\]

- For $\mu = 1$

\[
(A + \mathbf{I} \mu \mathbf{U}^t \mathbf{U})^{-1} = A^{-1} \left[ A^{-1} - \frac{\mathbf{U} \mathbf{U}^t}{\mathbf{I} + \mu \mathbf{U}^t A^{-1} \mathbf{U}} \right] A^{-1}
\]

Hence apply this to get the inverse of $\mathbf{Z} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{pmatrix}$

For what values of $\mu$ is $\mathbf{Z} > 0$?

<table>
<thead>
<tr>
<th>Lower Right Corner of $\mathbf{H} = \mathbf{Z}$</th>
<th>can be expressed as</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbf{G}_{22}$</td>
<td>$\mathbf{G}<em>{22} - \mathbf{G}</em>{21} \mathbf{G}<em>{11}^{-1} \mathbf{G}</em>{12}$</td>
</tr>
</tbody>
</table>
But now suppose $\mathcal{L}$ is a grammar, rank $\mathcal{L} = p$.

\[
\mathcal{L} = X' X \\
\text{where } X = (X_1, X_2) \\
\left( \mathcal{L}_{22} = X_2^t X_2 \right)
\]

and

\[
X = (X_1, X_2) \\
\left\{ \begin{array}{l}
\hat{X}_2 = \frac{1}{\epsilon_{21}} (X_1) \cdot X_2 \\
\hat{X}_2 = \frac{1}{\epsilon_{22}} (X_1) \cdot X_2
\end{array} \right.
\]

The

\[
\mathcal{L}_{22} - \mathcal{L}_{22} \mathcal{L}_{12} \mathcal{L}_{22} = \frac{1}{\epsilon_{22}} X_2^t X_2 = \mathcal{L}_{22}^{-1} \quad \text{(page A2.11)}
\]

57

\[
(\mathcal{L}^{-1})_{22} = \frac{1}{\epsilon_{22}} \mathcal{L}_{22}^{-1} \\
(\mathbf{X}^t \mathbf{X})_{22} = (\mathbf{X}_2^t \mathbf{X}_2)^t
\]

"The lower corner of the inverse grammar is the inverse of the ortho-grammar."

Hand 10. Show that $\mathcal{L}_{22}^{-1} = X_2^t X_2$, and relate this to property $\mathbf{x}^t \mathbf{x} = \mathbf{y}^t \mathbf{y}$. 

\[
\frac{1}{2} \mathbf{x}^t \mathbf{x} = \frac{1}{2} \mathbf{y}^t \mathbf{y}.
\]
Statistical Interpretation

If random vector $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$ has

 covariance matrix $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, then

$$\mathbb{E} \left( \mathbf{Y} - \Sigma^{-1} \begin{pmatrix} \Sigma_{12} \\ \Sigma_{22} \end{pmatrix} \right) \Sigma^{-1} \begin{pmatrix} \Sigma_{12} \\ \Sigma_{22} \end{pmatrix} \mathbb{E} \left( \mathbf{Y} - \Sigma^{-1} \begin{pmatrix} \Sigma_{12} \\ \Sigma_{22} \end{pmatrix} \right)^\top$$

where

$$\Sigma^{-1} \begin{pmatrix} \Sigma_{12} \\ \Sigma_{22} \end{pmatrix} = \Sigma_{12} - \Sigma_{22} \Sigma_{22} \Sigma_{22}^{-1}$$

is the covariance matrix of

$$\mathbf{U} = \mathbf{Y} - \Sigma^{-1} \begin{pmatrix} \Sigma_{12} \\ \Sigma_{22} \end{pmatrix} \mathbb{E} \left( \mathbf{Y} - \Sigma^{-1} \begin{pmatrix} \Sigma_{12} \\ \Sigma_{22} \end{pmatrix} \right)^\top$$

the residual of $Y_2$ after linear regression on $Y_1$;

that is, it's the "partial covariance of $Y_2$ given $Y_1$".

Partial Correlation

If $\Sigma_{12} = 0$ then $Y_1$ and $Y_2$ are uncorrelated.

If $\Sigma_{12} = 0$ then $Y_1$ and $Y_2$ are uncorrelated.
have zero partial correlations [after regression on other p - 2 covariates]

Hence II / verify this last statement.

Fisher Information

Observe \( X = f_\theta(x) \), \( \theta \in \mathbb{R}^p \), \( \theta = (\theta_1, \ldots, \theta_p) \)

Score Functions

\[ U_i = \frac{2}{\sigma_x^2} \log f_\theta(x) \quad (U_1, U_2) \]

Information Matrix

\[ \text{cov}_\theta(U) = \begin{pmatrix} d_{\theta 11} & d_{\theta 12} \\ d_{\theta 21} & d_{\theta 22} \end{pmatrix} \]

MLE

\( \hat{\theta} \rightarrow N_p (\theta, \frac{d_\theta^{-1}}{\sigma_x^2}) \), so

\[ \text{cov} (\hat{\theta}^2) = \left( \frac{d_\theta^{-1}}{\sigma_x^2} \right)_{22} = \left( \frac{d_{\theta 22} - d_{\theta 21} \theta_1 d_{\theta 11}}{\sigma_x^2} \right). \]

But if \( \theta_1 \) is known

\[ \text{cov} (\hat{\theta}_2) = \left( d_{\theta 22} \right) \]

So if \( \theta_1 \) is unknown, one loses Fisher information and gain's covariance according to the amount of projection of \( U_2 \) and \( U_1 \).
Dear Rohit,

Here is a first analysis of the itmn data you sent me last week.

[1] My notation:

\[ \text{Tt} = \text{treatment}, \, 1,2,\ldots,6 \]
\[ \text{Hour} = \text{time of visit}, \, 0,2,4,\ldots,360 \]
\[ y = \log_{10}(\text{RNA}), \text{the response variable.} \]

[2] I made the \( 46 \times 17 \) matrix "M", with \( y \) displayed versus patient number 1,2,\ldots,46 and visit number 1,2,\ldots,17. There were only a few missing values, which I filled in by the patient average.

[3] Letting \( M_0 \) be the \( 46 \times 16 \) matrix \( V \) but with the first column (the baseline measurement) removed, I computed the singular value decomposition

\[ M_0 = U \times \times d \times \times V', \]

where \( d \) is the diagonal matrix of singular values (square roots of the eigenvalues of \( V_0' \times \times V_0 \)), \( U \) is the matrix of left eigenvectors, and \( V \) is the matrix of right eigenvectors. Most of the variability was explained by the first 3 components, so \( U \) in * is \( 46 \times 3 \), \( V=16 \times 3 \), and \( d \) is \( 3 \times 3 \).

[4] the upper left panel of figure 4 shows the first three right eigenvectors \( v_1 \), \( v_2 \), \( v_3 \) (the columns of \( V \)): \( v_1 \) is just about the mean of 16 time measurements, \( v_2 \) is proportional to the rate of early decline in \( y \), and \( v_3 \) measures the late resurgence of \( y \), after about Hour 150. The proportion of \( M_0 \) explained by the 3 eigenvectors is 73% for \( v_1 \), 11% for \( v_2 \), and 5% for \( v_3 \).

[5] The "Loading Matrix" \( L=U\times\times d \) measures inner products, with

\[ L[i,j] = \text{inner product of patient } i \text{'s } y \text{ vector with the } \]
\[ \text{ith eigenvector } v_i \]

(not including the baseline measurement.)

The upper right panel of figure 4 plots \( L[i,2] \) versus \( L[i,1] \), with the plotting symbol indicating the Treatment group. Both components, the mean and the early decline, are seen to be useful for discriminating between the combination and monotherapies: groups 5 and 6 tend to have high values on both components, while groups 2 and 3 are generally lower. (Patient 39, in group 6, is seen be atypically low for both loadings.)

[6] The lower left panel of figure 4 plots \( L[i,3] \) versus \( L[i,2] \), i.e. the late rise versus the early decline. We see that \( L[i,3] \) tends to have high values in groups 2 and 3, these being the ones that allowed late resurgence in log RNA levels. Patient 39 is NOT atypical of Trt 6 on this last measure. The lower right panel, \( L[i,3] \) vs \( L[i,1] \) looks much the same.

[7] In this case I've abstracted Three summary statistics for each patient, \( L[i,1], L[i,2], L[i,3] \). There's a lot of ways one could put these together into a single test statistic, but probably the simplest is to compute

\[ S[i] = L[i,1] + L[i,2] - L[i,3]. \]
Here the minus sign comes from the fact that large values of \( L[i,3] \) indicate poor results, the opposite of \( L[i,1] \) and \( L[i,2] \). Figure 5 plots \( S[i] \) as a function of the Treatment group, and it looks like \( S \) is doing a good job of separating the combined regimens (triangles) from the monotherapies (diamonds.)

[8] A Wilcoxon test comparing Trt 2 with Trt 5 gave two-sided p-value .005 against the null hypothesis. Similarly, Trt 3 vs Trt 6 gave Wilcoxon p-value .003.

[9] This analysis is "honest" in the sense that the choice of Loading statistics didn't depend on the Treatment labels: if I permuted all 46 labels, I would get the Loading numbers for each patient. To put it another way, a full permutation analysis of the results above would yield the same p-values.

Hope this is helpful, let me know about questions,

Brad
fig4. First 3 principal vectors for vy 46x16, (baseline removed) I'mn data; 1=73%, 2=11%, 3=5%; 11/15/08

Now loadings 2 and 3

Now loadings 1 and 3
fig5. Combined score load1+load2-load3 for itmn data plotted against Treatment. 11/15/08