Math 6266 Linear Statistical Models **Fall 2019**

Lecture Notes

Instructor: Prof. Vladimir Koltchinskii Student: Yuanzhe Ma (yma412@gatech.edu)

Contents

1 Regression Problems

Given random (X,Y) where $X \in S$ and $Y \in \mathbb{R}$, our goal is to approximate Y by a function $g(X)$.

$\text{MSE}(g) := \mathbb{E}(Y-g(X))^2$, optimal $g_* = \text{argmin}_{g:S\to\mathbb{R}}\text{MSE}(g)$ where g is a measurable function.

Solution:

Assume $\mathbb{E}(Y^2) < \infty$, $g_*(X) = \mathbb{E}(Y|X)$, or $g_*(x) = \mathbb{E}(Y|X=x)$.

Proof

For any $g : S \to \mathbb{R}$, we have

$$\mathbb{E}(Y-g(X))^2 = \mathbb{E}(Y-g_*(X)+g_*(X)-g(X))^2$$

$$= \mathbb{E}(Y-g_*(X))^2 + 2\mathbb{E}(Y-g_*(X))(g_*(X)-g(X)) + \mathbb{E}(g_*(X)-g(X))^2$$

Note that

$$\mathbb{E}(Y-g_*(X))(g_*(X)-g(X)) = \mathbb{E}(\mathbb{E}[(Y-g_*(X))(g_*(X)-g(X))|X])$$

When X is fixed, $(g_*(X)-g(X)$ is a constant and $E((Y-g_*(X))$ given $X=x$ is 0, so $\mathbb{E}(Y-g_*(X))(g_*(X)-g(X)) = 0$.

Therefore,

$$\mathbb{E}(Y-g(X))^2 = \mathbb{E}(Y-g_*(X))^2 + \mathbb{E}(g_*(X)-g(X))^2 \geq \mathbb{E}(Y-g_*(X))^2$$

□

Moreover, if $\mathbb{E}(Y-g(X))^2 = \mathbb{E}(Y-g_*(X))^2$, then $\mathbb{E}(g(X)-g_*(X))^2 = 0 \implies g(X) = g_*(X)$ with probability 1.

Definition 1.1. Regression Function

$g_*(x) := \mathbb{E}(Y|X=x)$ is the regression function.

Regression in Statistics:

Given n iid data (X_i,Y_i), goal is to estimate $g_*(x)$ based on (X_i,Y_i).

Definition 1.2. Least Square Estimator

Let $\mathscr{G}$ be a set of function $g : S \to \mathbb{R}$ such that either $g_* \in \mathscr{G}$, or g_* has a reasonable approximation by the functions from $\mathscr{G}$. Define

$$\hat{g} := \text{argmin}_{g\in\mathscr{G}}\frac{1}{n}\sum_{i=1}^{n}(Y_i-g(X_i))^2$$

A choice of : $h_1, \cdots, h_N : S \to \mathbb{R}$ (a dictionary).

$$\mathscr{G} := \text{linear span}(\{h_1, \cdots, h_N\}) = \{\sum_{j=1}^{N} c_j h_j ; c_j \in \mathbb{R}, j = 1, 2, \cdots, n\}$$

So $\mathscr{G}$ is a linear space with dimension $\leq N$.

Example 1.1. $S = \mathbb{R}$, dictionary $(1, x, x^2, x^3, \cdots, x^k)$, so $\mathscr{G}$ is the true space of all polynomials of degree $\leq k$.

If $\mathbf{Y} = \begin{bmatrix} Y_1 \\ \cdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n$ is the response vector, then

$\forall g \in \Leftrightarrow g = \sum_{j=1}^{N} c_j h_j, \mathbf{c} = \begin{bmatrix} c_1 \\ \cdots \\ c_N \end{bmatrix} \in \mathbb{R}^N$,

$$\begin{bmatrix} g(X_1) \\ \cdots \\ g(X_n) \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{N} c_j h_j(X_1) \\ \cdots \\ \sum_{j=1}^{N} c_j h_j(X_n) \end{bmatrix} = \mathbf{Ac}$$

where the design matrix $A := (h_j(X_i))_{i=1,\cdots,n; j=1,c\cdots,N}$ is a $n \times N$ matrix.

Least Square $\Leftrightarrow \hat{\mathbf{c}} := \text{argmin}_{\mathbf{c} \in \mathbb{R}^N} \|\mathbf{Y} - \mathbf{Ac}\|^2$ and $\hat{g} = \sum_{j=1}^{n} \hat{c}_j h_j$.

Regression Model:

Given random (X, Y), $Y = g_*(X) + \xi$ where ξ is random noise.

Assumptions:

1) X and ξ are independent random variables.

2) $\mathbb{E}\xi = 0, \mathbb{E}\xi^2 = \sigma^2 < \infty$.

So $\mathbb{E}(Y|X) = g_*(X)$. ($g_*(X)$ is the regression function).

(X_i, Y_i) iid, $Y_j = g_*(X_j) + \xi_j$ and ξ_j iid.

Conditionally on X_j, we can view this regression model as a model with fixed (non-random) design.

Suppose $g_* \in=$ linear span$(\{h_1, \cdots, h_N\})$, $g = \sum_{j=1}^{N} c_j^* h_j, \mathbf{c}^* = \begin{bmatrix} c_1^* \\ \cdots \\ c_N^* \end{bmatrix} \in \mathbb{R}^N$.

$Y_i = \sum_{j=1}^{N} c_j^* h_j(X_i) + \xi_i$ and the design matrix $A := (h_j(X_i))_{i=1,\cdots,n; j=1,c\cdots,N}$ is a $n \times N$ matrix.

$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \cdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n$, $\xi := \begin{bmatrix} \xi_1 \\ \cdots \\ \xi_n \end{bmatrix} \in \mathbb{R}^n$ is the noise vector.

$\mathbf{Y} = \mathbf{Ac}^* + \xi$ is called linear regression model.

General Linear Regression (GLM): $Y = \mathbf{X}\beta^* + \xi$ with unknown variance of the noise.

2 Linear Algebgra

Definition 2.1. Minkowski sum

Suppose V is a vector space (linear space), $C_1, \cdots, C_k \subset V$, define the Minkowski sum as $C_1 + \cdots + C_k := \{x_1 + \cdots + x_k : x_j \in C_j\}$.

If $L_1, \cdots, L_k$ are subspaces of V, then their Minkowski sum $L_1 + \cdots + L_k = \text{linear span}(L_1 \cup \cdots \cup L_k)$ is also a subspace of V. Note that $L_1 \cup \cdots \cup L_k$ is not a linear space since it does not contain all linear combinations in it, but $\text{linear span}(L_1 \cup \cdots \cup L_k)$ is a linear space and it's larger than $L_1 \cup \cdots \cup L_k$.

Definition 2.2. Direct sum

$L = L_1 \oplus L_2 \oplus \cdots \oplus L_k$ is the direct sum of $L_1, \cdots, L_k$ if and only if for any $x \in L$, there exists unique $x_1 \in L_1, \cdots, x_k \in L_k$ such that $x = x_1 + \cdots + x_k$.

$$L = L_1 \oplus L_2 \oplus \cdots \oplus L_k \Leftrightarrow [0 = x_1 + \cdots + x_k, x_j \in L_j \implies x_j = 0, \forall j]$$

Proposition 2.1. If $L_1, L_2 \subset V$, then $\dim(L_1 + L_2) = \dim(L_1) + \dim(L_2) - \dim(L_1 \cap L_2)$

Proof

Choose a basis $l_1, \cdots, l_m$ of $L_1 \cap L_2$, $m = \dim(L_1 \cap L_2)$, extend this basis to the basis $l_1, \cdots, l_m, f_1, \cdots, f_l$ of L_1.

Extend the same basis to the basis of $L_2 : l_1, \cdots, l_m, g_1, \cdots g_k$.

Need to prove that $l_1, \cdots, l_m, f_1, \cdots, f_l, g_1, \cdots, g_k$ is the basis of $L_1 + L_2$.

$\dim(L_1 \cap L_2) = m$

$\dim(L_1) = m + l$

$\dim(L_2) = m + k$

$\dim(L_1 + L_2) = m + l + k$ $\square$

So

$$L = L_1 \oplus L_2 \Leftrightarrow L_1 \cap L_2 = \{0\} \Leftrightarrow \dim(L_1 \cap L_2) = 0 \Leftrightarrow \dim(L_1 + L_2) = \dim(L_1) + \dim(L_2)$$

Proposition 2.2. Suppose $L_1, \cdots, L_k$ are subspaces of V and $L = L_1 + \cdots + L_k$, then the following statements are equivalent:

(i) $L = L_1 \oplus \cdots \oplus L_k$

(ii) $\forall i = 1, \cdots, k-1, L_i \cap (L_{i+1} + \cdots + L_k) = \{0\}$

(iii) $\dim(L) = \dim(L_1) + \cdots + \dim(L_k)$

Example 2.1. If $L_1, \cdots, L_k$ are linear spaces, define the Cartesian product operation as follows: $L_1 \oplus \cdots \oplus L_k := \{(x_1, \cdots, x_k) : x_j \in L_j\}$.

Note that $(x_1, \cdots, x_k) + (x_1^{'}, \cdots, x_k^{'}) = (x_1 + x_1^{'}, \cdots, x_k + x_k^{'})$. Then let $L_j^{'} = \{(0, \cdots, 0, x, 0, \cdots, 0), x \in L_j\}$ where x is in the j th position, then it's a subspace of $L_1 \oplus \cdots \oplus L_k$.

In addition, $\underbrace{L_1^{'} \oplus \cdots \oplus L_k^{'}}_{\text{the usual direct sum}} = \underbrace{L_1 \oplus \cdots \oplus L_k}_{\text{the Cartesian product we just defined}}$.

Theorem 2.1. *Projection Theorem*

Suppose $(v, \langle ., . \rangle)$ is an inner product space (A inner product space which is complete is called a ***Hilbert space****), and $C \subset V$ is a closed convex set, then for all $x \in V$, there exists a unique $P_C(x) \in C$ such that*

$$\|x - P_C x\| = \inf_{y \in C} \|x - y\|$$

Proof

Define $\Delta := \inf_{y \in C} \|x - y\|$, then there exists a sequence $\{y_n\}, y_n \in C$ such that $\|x - y_n\| \to \Delta$ as $n \to \infty$.

Also,

$$\Delta \leq \|x - \frac{y_n + y_m}{2}\| = \frac{1}{2}(\|x - y_n\| + \|x - y_m\|) \to \Delta$$

So

$$\Longrightarrow \|x - \frac{y_n + y_m}{2}\| \to \Delta$$

By the parrallelogram identity $(\|u + v\|^2 + \|u - v\|^2 = 2(\|u\|^2 + \|v\|^2))$, take $u = x - y_n$ and $v = x - y_m$

We have $\|u\|^2 \to \Delta^2$, $\|V\|^2 \to \Delta^2$, $\|u + V\|^2 \to 4\Delta^2$

$$\|y_n - y_m\| \to 0 \Longrightarrow \exists \lim_n y_n := P_C x \in C$$

$$\|x - P_C x\| = \lim_{n\to\infty} \|x - y_n\| = \inf_{y\in C} \|x - y\|$$

Since $y \to \|x - y\|^2$ is strictly convex (convex function: $f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2)$ for any $x_1 \ne x_2, \lambda \in (0,1)$), the minimum is unique. □

Definition 2.3. Orthogonal projection onto an affine subspace

Let S be a subspace of a finite dimensional inner product space V and $A = a + S$ be an affine subspace with parallel space S. The orthogonal projection $P_A : V \to A$ onto A is defined by $P_A(v) = a + P_S(v - a)$ where P_S is the corresponding orthogonal projection onto S.

Definition 2.4. Orthogonal Complement

Let $L \subset V$ be a subspace, define its orthogonal complement as $L^\perp = \{u \in V : u \perp L\}$.

Proposition 2.3. For any $x \in V$, there exists a unique vector $\hat{x} \in L$, such that $x - \hat{x} \in L^\perp$. Moreover, $\hat{x} = P_L x$.

Proof

Take $\hat{x} = P_L x$, we want to show that $x - \hat{x} \perp L$.

Suppose not! Then there exists a $h \in L$ such that $\langle x - \hat{x}, h\rangle \ne 0$ where $h \ne 0$.

Without loss of generality, assume that $\langle x - \hat{x}, h\rangle > 0$, if we can make $\|x - \hat{x}\|^2$ smaller, then we can get the contradiction.

To see this, note that for some small $t > 0$, we have

$$\|x - (\hat{x} + th)\|^2 = \|x - \hat{x}\|^2 \underbrace{-2t\langle x - \hat{x}, h\rangle}_{<0} + t^2\|h\|^2$$

For t is small, $t >> t^2$, so $\|x - (\hat{x} + th)\|^2 < \|x - \hat{x}\|^2$.

So $P_L x$ is the desired vector. Furthermore, for any $y = \hat{x} + a, a \in L, \langle x - y, l\rangle$ does not always equal to 0 for all $l \in L$, so $P_L x$ is the unique one.

□

Definition 2.5. self-adjoint operator

Suppose $(V, \langle ., .\rangle)$ is a inner product space and $A : V \to V$ is a linear operator (transformation).

We say A is self-adjoint if $\langle Ax, y\rangle = \langle x, Ay\rangle, x, y \in V$. (in matrix space self-adjoint is equivalent to Hermitian matrix).

Definition 2.6. range and kernel of a subspace

$Im(A) = R(A) : \{Ax, x \in V\} \subset V$ is a subspace of V.

$Ker(A) = n(A) = \{x : Ax = 0\} \subset V$ is a subspace of V.

Recall a previous proposition: if $L \subset V$ is a subspace of V, then for any $x \in V$, there exists a unique $P_L x$ such that $||x - P_L x|| = \inf_{y\in L} ||x - y||$. Moreover, $P_L x$ is uniquely characterized by the following relationship

(1) $x - P_L x \in L^{\perp}$

(2) $P_L x \in L$

Theorem 2.2. *1. Suppose that $e_1, \cdots, e_k \in L$ are orthonormal bases ($\langle e_i, e_j\rangle = \delta_{ij}$) of the subspace L,* linear span$(e_1, \cdots, e_k)$

$= L$, then $P_L x = \sum_{j=1}^{k} \langle x, e_j\rangle e_j$.

2. Suppose that $e_1, \cdots, e_k \in L$ are orthogonal bases ($\langle e_i, e_j\rangle = \delta_{ij} ||e_j||^2$) of the subspace L, linear span$(e_1, \cdots, e_k) = L$, *then $P_L x = \sum_{j=1}^{k} \frac{\langle x, e_j\rangle}{\langle e_j, e_j\rangle} e_j$.*

3. If matrix P projects a vector into the column space of A, then $P = A(A^T A)^{-1} A^T$.

Proof

Need to show:

(1) $x - P_L x \in L^{\perp}$ which is equivalent to $x - P_L x \perp e_j, j = 1, \cdots, k$

$\langle x - P_L x, e_j\rangle = \langle x, e_j\rangle - \langle P_L x, e_j\rangle = \langle x, e_j\rangle - \sum_{i=1}^{k} \langle x, e_i\rangle\langle e_i, e_j\rangle = \langle x, e_j\rangle - \sum_{i=1}^{k} \langle x, e_i\rangle \delta_{ij} = \langle x, e_j\rangle - \langle x, e_j\rangle = 0$

(2) $P_L x \in L$, obvious. □

Proposition 2.4. For an orthogonal projection $P_L : V \to V$, the following properties hold (conversely also true, the following properties indicate it is an orthogonal projection) :

(i) P_L is a linear operator, $P_L(x + y) = P_L(x) + P_L(y)$.

(ii) P_L is self-adjoint.

Proof: $\langle P_L x, y\rangle = \langle P_L x, P_L y + P_{L^{\perp}} y\rangle = \langle P_L x, P_L y\rangle = \langle P_L x + P_{L\perp} x, P_L y\rangle = \langle x, P_L y\rangle$.

(iii) $P_L^2 = P_L$ (idempotent).

(iv) $Im(P_L) = L$, $Ker(P_L) = L^{\perp}$.

Proposition 2.5. Suppose $A : V \to V$ is a linear self-adjoint operator and $A^2 = A$, then $A = P_L$ where $L = Im(A)$.

Proof

Clearly, for any $x \in V$, $Ax \in L$, it's sufficient to check that $x - Ax \perp L$.

For any $y \in V$, we need $\langle x-Ax, Ay\rangle = \langle x, Ay\rangle - \langle Ax, Ay\rangle = \langle x, Ay\rangle - \langle x, A^2y\rangle(\mathrm{self-adjoint}) = \langle x, Ay\rangle - \langle x, Ay\rangle(\mathrm{idempotent}) = 0$. □

Proposition 2.6. Suppose $P_1, \cdots, P_k$ are orthogonal projections in V, say $P_j = P_{L_j}$, and let $P = P_1 + \cdots + P_k$, then the following statements are equivalent:

(i) P itself is an orthogonal projection.

(ii) $P_iP_j = 0$ when $i \neq j$.

(iii) $L_i \perp L_j$ when $i \neq j$.

(iv) $P = P_L$ where $L = L_1 \oplus \cdots \oplus L_k$.

Proof

(i) to (ii): for any $x \in V$, $||x||^2 \geq ||Px||^2 = \langle Px, Px\rangle = \langle P^2x, x\rangle = \langle Px, x\rangle = \langle \sum_{j=1}^k P_jx, x\rangle = \sum_{j=1}^k \langle P_j^2 x, x\rangle = \sum_{j=1}^k \langle P_jx, P_jx\rangle = \sum_{j=1}^k ||P_jx||^2$.

For $x = P_jy, y \in V$, we have $||P_iy||^2 \geq \sum_{i=1}^k ||P_jP_iy||^2 = ||P_iy||^2 + \sum_{j\neq i} ||P_iP_jy||^2$ so $P_jP_jy = 0$ for $y \in V, j \neq i$, which means $P_jP_i = 0$ for $j \neq i$.

(ii) and (iii) are equivalent:

$P_iP_j = 0 \Leftrightarrow \forall y \in V \ \ P_iP_jy = 0 \Leftrightarrow \forall y \in V \ \ P_jy \in Ker(P_i) = L_i^{\perp}$.

This implies that $L_j = Im(P_j) \subset L_i^{\perp} \implies L_j \subset L_i^{\perp}, i \neq j \Leftrightarrow L_i \perp L_j$.

(iii) to (iv): Need to check P is an orthogonal projection, since P is self-adjoint, enough to check $P^2 = P$.

$P^2 = \sum_{i=1} P_i^2 + \sum_{i\neq j} P_iP_j = \sum_{i=1} P_i^2 = \sum_{i=1} P_i = P$.

$Im(P) = L_1 \oplus \cdots \oplus L_k$. (direct sum is immediate because we are under assumptions that that $L_i \perp L_j$ so it will be direct sum, which means we can have a unique representation) □

Corollary 2.3. *Let I be identity operator ($Ix = x$ thus an orthogonal projection) and $P_1, \cdots, P_k$ are orthogonal projections in V, say $P_1, \cdots, P_k$ is a resolution (split) of identity, $P_1 + \cdots + P_k = I$.*

We have the following properties:

$P_i P_j = 0, i \neq j.$

$P_i = P_{L_i} \implies L_i \perp L_j$.

$V = L_1 \oplus \cdots \oplus L_k.$

Theorem 2.4. *Algebraic form of Cochran Theorem: Suppose $T_1, \cdots, T_k$ are self-adjoint linear operators in V with $Im(T_j) = L_j$, let $P = T_1 + \cdots + T_k$ (sum of operators), and P is an orthogonal projection, say $P = P_L, L \subset V$, then the following four statements are equivalent:*

(i) For any i, T_i is an orthogonal projection, in other words, $T_i = T_i^2$.

(ii) $L_i \perp L_j$ if $i \neq j$, or $L = L_1 \oplus \cdots \oplus L_k$.

(iii) $\dim(L) = \dim(L_1) + \cdots + \dim(L_k)$. *(commonly used condition)*

(iv) $T_i T_j = 0$ if $i \neq j$.

Proof

Suppose $P = I, L = V$, otherwise we can define $T_{k+1} = P_{L^\perp}$ and the $T_1 + \cdots T_{k+1} = P + T_{k+1}$ is the identity operator.

(i) to (ii): See previous proposition (sum of T_i is an orthogonal projection, so it's obvious).

(ii) to (iii): $L_1 \oplus \cdots \oplus L_k \implies \dim(L) = \dim(L_1) + \cdots + \dim(L_k)$, obvious.

(iii) to (iv): $P = I$, we can write $x \in V, x = Ix = T_1 x(\in L_1) + \cdots + T_k x(\in L_k) \implies V = L_1 + \cdots + L_k$. In addition, $\dim(L) = \dim(L_1) + \cdots + \dim(L_k)$ from previous proposition, so $V = L_1 \oplus \cdots \oplus L_k$. So such representation is unique.

Take $x = T_i y, y \in V$, so $T_i y = \sum_{j=1}^n T_i T_j y = T_i y + \sum_{j \neq i} T_j T_i y$.

$\implies T_i y = T_i^2 y, T_j T_i y = 0 (i \neq j) \implies T_j T_i = 0.$

(iv) to (i): Enough to prove $T_i = T_i^2$, $T_i - T_i^2 = T_i(I - T_i) = T_i \sum_{j \neq i} T_j = \sum_{j \neq i} T_i T_j = 0 \implies T_i = T_i^2$. □

Proposition 2.7. Given $A : V \to V$ is a finite-dimension linear operator, let $A \subset V$, we say L is an invariant subspace of A if $A(L) = \{Ax, x \in L\} \subset L$. If A is self-adjoint and $L \subset V$ is an invariant subspace, then $L^\perp$ is also an invariant subspace.

Proof

Need to prove that for $x \in L^{\perp}, Ax \in L^{\perp}$, or $\langle Ax, \underbrace{y}_{\in L} \rangle = 0$ for all $y \in L$.

Note that $\langle Ax, y \rangle = \langle x, \underbrace{Ay}_{\in L} \rangle = 0$. □

Theorem 2.5. *Spectral theorem for self-adjoint operator*

Let $A : V \to V$ be a self-adjoint linear operator, then there exists a finite set $S \subset \mathbb{R}$ and a resolution of I (split of identity operator as sum of orthogonal operators $\{P_\lambda, \lambda \in S\}$, i.e. $\sum_\lambda P_\lambda = I, P_\lambda P_{\lambda'} = 0$ for $\lambda \neq \lambda'$ and $Im(P_\lambda) \perp Im(P_{\lambda'})$) such that $A = \sum_{\lambda \in S} \lambda P_\lambda$. Moreover, $S = \sigma(A)$ (the set of eigenvalues, might not all be distinct) and for any $\lambda \in \sigma(A), P_\lambda = P_{L_\lambda}$ where L_λ is the eigenspace of A for eigenvalue λ. P_λ are called spectral projections of our operator A.

Proof

$f_A(x) = \langle Ax, x \rangle$ is the quadratic form of A (maps from $V \to \mathbb{R}$), and it's clearly continuous for any finite dimensional space, consider $\{x : ||x|| = 1\}$ (a compact set so attains max and min).

Define $e_1 := \text{argmax}_{||x|=1} \langle Ax, x \rangle$ and $\lambda_1 := \max_{||x|=1} \langle Ax, x \rangle$ and $L_1 = \text{linear span}(e_1)$, and we will prove that $Ae_1 = \lambda_1 e_1$.

We can write $Ae_1 = \lambda_1 e_1 + h$ for some vector h, need to show $h = 0$. $Ae_1 = \lambda_1 e_1 + h =$

$\underbrace{\langle Ae_1, e_1 \rangle e_1}_{\in P_{L_1}(Ae_1) \text{ because } L_1 \text{ is spanned by } e_1} + h$, the residual of a orthogonal projection should be zero, so $h \perp L_1$.

Assume $h \neq 0$, let $v = \frac{e_1 + th}{||e_1 + th||}$ (t is small and positive), then we need to show that there exists a t such that $\langle Av, v \rangle > \langle Ae_1, e_1 \rangle = \lambda$, leading to a contradiction.

Note that $\langle Ae_1, h \rangle = \langle \lambda_1 e_1 + h, h \rangle = ||h||^2$

$$\langle Av, v \rangle = \frac{\langle A(e_1 + th), e_1 + th \rangle}{\langle e_1 + th, e_1 + th \rangle} = \frac{\langle Ae_1, e_1 \rangle + 2t\langle Ae_1, e_1 \rangle + t^2 \langle Ah, h \rangle}{1 + t^2 ||h||^2} = \frac{\lambda_1 + ||h||^2 + 2t\langle Ah, h \rangle}{1 + t^2 ||h||^2}$$

Plus, if we want $\frac{\lambda_1 + ||h||^2 + 2t\langle Ah, h \rangle}{1 + t^2 ||h||^2} > \lambda_1 \Leftrightarrow \lambda_1 + ||h||^2 + 2t\langle Ah, h \rangle > \lambda_1(1 + t^2||h||^2) \Leftrightarrow 2t||h||^2 > (\lambda_1 ||h||^2 - \langle Ah, h \rangle)t^2$

It follows for a positive t which is small enough, $\langle Av, v \rangle > \lambda_1$, which is contradiction. Therefore, $h = 0$ and λ_1 is an eigenvalue and e_1 is an eigenvector.

Consequently, $L_1 = \text{linear span}(e_1)$ is an invariant subspace of A (eigenvectors, so map L_1 to L_1). Since A is self-adjoint, by the previous proposition, $L_1^{\perp}$ is also an invariant subsapce of A. Define $P_1 = P_{L_1}$ and let $A_1 = A - \lambda_1 P_1$, then $A_1 = 0$ on L_1 and $A_1 = A$ on $L_1^{\perp}$ (minus something irrelevant). Moreover, $A_1 : L_1^{\perp} \to L_1^{\perp}$ ($\dim L_1 \perp = d - 1$) is a self-adjoint operator (this comes from the fact that A is self-adjoint). It means that we can continue proof again (replace A_1 with A) in the first proof.

Do the previous proof again, define $e_2 := \text{argmax}_{||x|=1, x \in L_1} \langle Ax, x \rangle$ and $\lambda_2 := \max_{||x|=1, x \in L_1} \langle Ax, x \rangle$ and $L_2 =$

linear span(e_2), so $Ae_2 = \lambda_2 e_2$, again eigenvector, repeat this process, $L_2 =$ linear span$(e_2), P_2 = P_{L_2}$. Let $A_2 = A - \lambda_1 P_1 - \lambda_2 P_2, A_2 : (L_1 \oplus L_2)^\perp \to (L_1 \oplus L_2)^\perp$ is self-adjoint and $(L_1 \oplus L_2)^\perp$ is invariant with dimension $d-2$.

Continue this process, and if we have the dimension of V to be d, we will get $A = \sum_{j=1}^{D} \lambda_j P_j$ where $P_j = P_{L_j}$ are orthogonal projections on some space L_j (1 dimension, linear span$(e_j), L_i \perp L_j$).

After d steps, we construct $\lambda_1 e_1, \lambda_2 e_2, \cdots, \lambda_d e_d$ such that

(i) $\lambda_j \in \mathbb{R}$.

(ii) $e_1, \cdots, e_d$ is an orthonormal basis.

(iii) $Ae_j = \lambda_j e_j$.

(iv) $A = \sum_{j=1}^{d} \lambda_j P_j$ where P_j is a projection to L_j(linear span(e_j)).

The matrix of A in the basis of $\{e_j\}$ will be $\langle Ae_i, e_j \rangle_{i,j} = \langle \lambda_i e_i, e_j \rangle = \lambda_i \delta_{ij} = \text{diag}(\lambda_1, \cdots, \lambda_d)$ and some of them can be equal. $\sigma(A) = \{\lambda_1, \cdots, \lambda_d\}$ (this is a list with no repitition), some of them could be repeated, so card$(A) \leq d$.

A fact: eigenvalue of multiplicity k of a real symmetric matrix has exactly k linearly independent eigenvector

If one of the eigenvalues has multiplicity k, we can choose k linearly independent eigenvectors, each with dimension 1, together with dimension k, so essentially, multiplicity is not a problem.

For any $\lambda \in \sigma(A)$ define $J_\lambda := \{j : \lambda_j = \lambda\}$, $P_\lambda = \sum_{j \in J_\lambda} P_j$.

So $A = \sum \lambda_j P_j = \sum_{\lambda \in \sigma(A)} \lambda P_\lambda$, where each P_λ has dimension $\#(J_\lambda)$ (number of multiplicity).

In addition, $\sum_{j=1}^{d} P_j = I \implies \sum_{\lambda \in \sigma(A)} P_\lambda = I$.

□

Clearly the spectral decomposition is not unique (essentially because of the multiplicity of eigenvalues). But the eigenspaces corresponding to each eigenvalue are fixed. So there is a unique decomposition in terms of eigenspaces and then any orthonormal basis of these eigenspaces can be chosen.

Definition 2.7. Similar Matrix

Suppose A and B are two square matrices of size n . Then A and B are similar if there exists a nonsingular matrix S of size n such that $A = S^{-1}BS$, then we can interpret A and B are the same linear transformation under different basis.

Corollary 2.6. *SVD*

Spectral theorem for self-adjoint operator, just like the polar decomposition for complex numbers, $z = |z|e^{i\theta}$, we can decompose a self-adjoint operator into product of some positive definite matrix and rotation matrix that preserves distance), we can get the singular value decomposition (SVD).

Definition 2.8. Bilinear Form

A bilinear form on a vector space V is a bilinear map $V \times V \to K$, where K is the field of scalars. In other words, a bilinear form is a function $V \times V \to K$ that is linear in each argument separately:

$B(u+v,w) = B(u,w) + B(v,w)$ and $B(\lambda,v) = \lambda B(u,v)$.

$B(u,v+w) = B(u,v) + B(u,w)$ and $B(u,\lambda v) = \lambda B(u,v)$.

Corollary 2.7. *Suppose $A : V \to V$ be a linear operator with operator norm $||A|| := \sup_{||x||=1} ||Ax|| = \sup_{||x||=1,||y||=1} |\langle Ax,y\rangle|$. (This is true since $|\langle Ax,y\rangle| \leq ||Ax|| \cdot ||y|| = ||Ax|| \implies \sup_{||x||=1,||y||=1} \langle Ax,y\rangle \leq \sup_{||x||=1} ||Ax||$ and $\sup_{||x||=1} ||Ax|| = \sup_{||x||=1} |\langle Ax, \frac{Ax}{||Ax||}\rangle| \leq \sup_{||x||=||y||=1} |\langle Ax,y\rangle|$). If A is self-adjoint, we have $||A|| = \sup_{\lambda\in\sigma(A)} |\lambda|$.*

Proof

Let eigenvectors $e_1,\cdots,e_d$ be the basis of A so $Ae_i = \lambda_i e_i$ and we can write $A = \mathrm{diag}(\lambda_1,\cdots,\lambda_d)$ in the new basis system.

We can write the bilinear form $\langle Ax,y\rangle = \sum_{j=1} \lambda_j x_j y_j$ (where $x_j = \langle x,e_j\rangle$ and $y_j = \langle y,e_j\rangle$) in the new basis system as well (in general $\langle Ax,y\rangle = \sum_{i,j} a_{ij} x_i y_j$).

So $|\langle Ax,y\rangle| = |\sum_{j=1} \lambda_j x_j y_j| \leq \sum_{j=1} |\lambda_j| \cdot |x_j| \cdot |y_j| \leq \max_j |\lambda_j| \sum_{j=1}^n |x_j||y_j| \leq \underbrace{\max_j |\lambda_j| (\sum_j x_j^2)^{\frac{1}{2}} (\sum_j y_j^2)^{\frac{1}{2}}}_{\text{Cauchy Schwarz}} = \max_j |\lambda_j| ||x|| \cdot ||y|| =$

$\max_j |\lambda_j|$ since we are given the norm is 1.

$||A|| \geq \max_j \langle Ae_j, e_j\rangle = \max_j |\lambda_j|$.

So $||A|| = \max_j |\lambda_j|$.

□

Corollary 2.8. *Define function of matrices based on spectral theorem: operator functional calculus.*

We can, for any function $f : \mathbb{R} \to \mathbb{R}$, define $f(A)$ where A is self-adjoint.

$A = \sum_\lambda \lambda P_\lambda$.

Then $A^2 = (\sum_\lambda \lambda P_\lambda)^2 = \sum_\lambda \lambda^2 P_\lambda$ since cross product is 0.

Similarly, $f(A) = \sum_\lambda f(\lambda) P_\lambda$ for any polynomial. Since any continuous function can be approximated by continuous functions, so we can define $f(A)$ for any continuous f, the domain of f can be pretty small as long as it contains all the λ.

Definition 2.9. adjoint operator

If we have $A : V_1 \to V_2$ and both are inner product spaces, there exsits a unique $A^* : V_2 \to V_1$, where $\langle Ax, y\rangle = \langle x, A^* y\rangle$ for all $x \in V_1, y \in V_2$. We call A^* be the adjoint operator of A. In matrix form, it's Hermitian matrix and they satisfy the following properties.

(i) $A^{**} = A$

(iii) $A^* = A \Leftrightarrow A$ is self-adjoint

(iii) $(A+B)^* = A^* + B^*$

(iv) $(AB)^* = B^* A^*$

(vi) If we have $A^*A : V_1 \to V_1$ and $AA^* : V_2 \to V_2$, they will be both self-adjoint and positive semi-definite. For example, $\langle A^*Ax, x\rangle = \langle Ax, Ax\rangle \geq 0$.

Definition 2.10. inverse

Let $A : V \to V$ be a linear opeartor, and $Ker(A) = \{x : Ax = 0\} = \{0\} \Leftrightarrow A$ is one to one.

In addition, $Im(A) = V \Leftrightarrow A$ is mapping onto V.

Then there exsits A^{-1} such that $AA^{-1} = A^{-1}A = I$, A is invertible and we call A^{-1} be the inverse of A.

Theorem 2.9. *Suppose $A : V_1 \to V_2$ is a linear operator, then there exists a unique operator (Moore-Penrose pseudoinverse) $A^+ : V_2 \to V_1$ such that*

(i) $AA^+A = A$

(ii) $A^+AA^+ = A^+$

(iii) $A^+A : V_1 \to V_1$ and $AA^+ : V_2 \to V_2$ are both self-adjoint. (the unique property for Moore-Penrose pseudoinverse).

If $V_1 = V_2 = V$ and A is invertible, then $A^+ = A^{-1}$, proof is obvious.

Proof

1: Prove uniqueness.

Assume there exists B such that properties (i) to (ii) hold, need to show $B = A^+$.

Define $C = AA^+ - AB$, then C is self-adjoint by (iii).

$C^2 = (AA^+ - AB)(AA^+ - AB) = AA^+AA^+ - ABAA^+ - AA^+AB + ABAB$

$= AA^+ - AA^+ - AB - AB = 0$

Since C is self-adjoint, we can use spectral theorem, we have $\lambda = 0$, it follows that $C = 0$. So $AA^+ = AB$.

Similarly, we have $A^+A = BA$.

$A^+ = A^+AA^+ = BAA^+ = BAB = B$. So $A^+ = B$ and unique.

2: Then prove existence.

Suppose operator $A^*A : V_1 \to V_1$ is invertible, then $A^+ = (A^*A)^{-1}A^*$.

(i) $AA^+A = A(A^*A)^{-1}A^*A = A$.

(ii) $A^+AA^+ = (A^*A)^{-1}A^*A(A^*A)^{-1}A^* = (A^*A)^{-1}A^* = A^+$.

(iii) $A^+A = (A^*A)^{-1}A^*A = I_{V_1}$ is self-adjoint, $(AA^+)^* = (A(A^*A)^{-1}A^*)^* = AA^+$ is self-adjoint.

Similarly, suppose $AA^* : V_2 \to V_2$ is invertible, then $A^+ = A^*(AA^*)^{-1}$ and $(AA^+)^* = (A(A^*A)^{-1}A^*)^* = A^{**}((A^*A)^{-1})^*A^* = A(A^*A)^{-1}A^* = AA^+$ so AA^+ is self-adjoint.

What if A^*A and AA^* are not invertible? We will use regularization, add a number positive times identity matrix to a matrix so that matrix can be inverted. Note that A^*A is positive semidefinte ($x^TA^*Ax = \langle A^*Ax, x\rangle = \langle Ax, Ax\rangle \geq 0$), and for any $t > 0$, $A^*A + tI_{v_1}$ is positive definite (all eigenvalues > 0 so invertible).

Proposition: There always exists $\lim_{t\to 0}(A^*A + tI_{V_1})^{-1}A^*$ and exists $\lim_{t\to 0}A^*(AA^* + tI_{V_2})^{-1}$, moreover, they equal to each other and are the unique A^+.

$$\lim_{t\to 0}(A^*A + tI_{V_1})^{-1}A^* = \lim_{t\to 0}A^*(AA^* + tI_{V_2})^{-1} = A^+$$

Proof

A^*A is self-adjoint, its spectral is $\sigma(A^*A) \subset \mathbb{R}^+$, and use spectral theorem we can get that $A^*A = \sum_{\lambda\in\sigma(A^*A)} \lambda P_\lambda$ (P_λ is projection onto eigenspace of λ) and $\{P_\lambda, \lambda \in \sigma(A^*A)\}$ forms a resolution of identity $I_{V_1} = \sum_{\lambda\in\sigma(A^*A)} P_\lambda$. $A^*A + tI_{V_1} = \sum_{\lambda\geq 0}(\lambda+t)P_\lambda$, so $(A^*A+tI_{V_1})^{-1} = \sum_{\lambda\geq 0}\frac{1}{\lambda+t}P_\lambda$.

Because $t > 0, (A*A+tI_{V_1})^{-1} = \sum_{\lambda\geq 0}\frac{1}{\lambda+t}P_\lambda$ exists.

$\lim_{t\to 0}(A^*A+tI_{V_1})^{-1}A^* = \lim_{t\to 0}\sum_{\lambda\geq 0}\frac{1}{\lambda+t}P_\lambda A^*$ and we are assuming that $0 \in \sigma(A^*A)$ so we have trouble calculating this limit.

If we want the limit to exist, our only hope is that for $\lambda = 0$, $P_0A^* = 0$ where P_0 is the projection onto $Ker(A^*A)$.

$P_0A^* = 0 \Leftrightarrow (P_0A^*)^* = 0 \Leftrightarrow AP_0 = 0$ (P_0 is projection hence self-adjoint).

Observe that $A^*AP_0 = 0$ because we can use the spectral theorem

$$A^*AP_0 = \sum_{\lambda\geq 0}\lambda P_\lambda P_0 = 0P_0^2 + \sum_{\lambda>0}\lambda P_\lambda P_0 = 0+0 = 0$$

since for $\lambda \neq 0, P_\lambda P_0 = 0$.

$P_0 = P_{L_0}$ where $L_0 = Ker(A^*A)$, for $x \in L_0$, $A^*Ax = 0 \implies \langle A^*Ax, x\rangle = 0 \implies \langle Ax, Ax\rangle = 0 \implies ||Ax|| = 0 \implies Ax = 0 \implies AP_0 = 0$.

$$\lim_{t\to 0}(A^*A+tI_{V_1})^{-1}A^* = \lim_{t\to 0}\sum_{\lambda\geq 0}\frac{1}{\lambda+t}P_\lambda A^* = \lim_{t\to 0}\sum_{\lambda>0}\frac{1}{\lambda+t}P_\lambda A^* = \sum_{\lambda>0}\frac{1}{\lambda}P_\lambda A^*$$

□

So $(AA^+)^* = \sum_{\lambda>0}(\frac{1}{\lambda}AP_\lambda A^*)^* = \sum_{\lambda>0}\frac{1}{\lambda}(AP_\lambda A^*)^* = \sum_{\lambda>0}\frac{1}{\lambda}AP_\lambda A^* = AA^+$ is which self-adjoint, but not necessarily a projection because the identity resolution we construct is only on V_1 but AA^+ acts on V_2.

$\square$

Least square problem: We have linear transformations $A : V_1 \to V_2$ want $Ax \approx y, y \in V_2$.

(LS): $\min ||Ax-y||^2$ with respect to $x \in V_1$.

We are trying to project y on to $Im(A) = \{Ax : x \in V\}$.

(i) If $\hat{x}$ solves problem then $A\hat{x} = P_{Im(A)}y$. Note that $\hat{x}$ is not necessarily unique, we can always add h such that $Ah = 0$.

(i) If there are 2 solutions $\hat{x}_1, \hat{x}_2$, then $\hat{x}_1 - \hat{x}_2 \in Ker(A)$. (Indeed $A\hat{x}_1 = P_{Im(A)}y$ and $A\hat{x}_2 = P_{Im(A)}y$ so $A\hat{x}_1 = A\hat{x}_2 \implies A(\hat{x}_1 - \hat{x}_2) = 0$.

Proposition 2.8. The set of all solutions $\text{Argmin}_{x\in V_1} ||Ax-y||^2$ the set of all solutions of problem of LS = $A^+y + Ker(A)$ where A^+ is Moore-Penrose pseudoinverse.

Proof

Enough to show that AA^+y is the projection $P_{Im(A)}y$, which is equivalent to show $y - AA^+y \perp Im(A)$. In other words, for any $x \in V_1, y - AA^+y \perp Ax$, or $\langle y - AA^+y, Ax\rangle = 0 \Leftrightarrow \langle y, Ax\rangle = \langle AA^+y, Ax\rangle \Leftrightarrow \langle A^*y, x\rangle = \langle A^*AA^+y, x\rangle \Leftrightarrow A^*y = A^*AA^+y, \forall y \in V_2 \Leftrightarrow A^* = A^*AA^+ \Leftrightarrow (A^*)^* = (A^*AA^+)^* \Leftrightarrow A = \underbrace{(AA^+)^*}_{\text{self-adjoint}} A \Leftrightarrow A = AA^+A$

which is the definition of Moore-Penrose pseudoinverse. $\square$

3 Probability

Random variables and covariance inner product spaces: $(V, \langle\cdot,\cdot\rangle)$ with finite finite space. Let X be a random variable with values in V, assume that $\mathbb{E}[\langle X, u\rangle]$ is finite for any $u \in V$ (equivalent to existence of moments).

$u \in V \mapsto \mathbb{E}\langle X, u\rangle \in \mathbb{R}$ is a linear function on V.

So there exists $\mathbb{E}[X] \in V$ such that $\langle \mathbb{E}X, u\rangle = \mathbb{E}\langle X, u\rangle, u \in V$ and we call $\mathbb{E}[X]$ the expecation of V.

- $\mathbb{E}[c_1X_1 + c_2X_2] = c_1\mathbb{E}[X_1] + c_2\mathbb{E}[X_2]$
- For any $T : V \to V_1$ where V_1 is an inner product space, $\mathbb{E}[TX] = T\mathbb{E}[X]$.
  For any u, $\langle \mathbb{E}[TX], u\rangle = \mathbb{E}\langle TX, u\rangle = \mathbb{E}\langle X, T^*u\rangle = \langle \mathbb{E}X, T^*u\rangle = \langle T\mathbb{E}[X], u\rangle \implies \mathbb{E}[TX] = T\mathbb{E}[X]$.

Proposition 3.1. If $B(u,v)$ is a bilinear form on V, then there exists a linear operator $B : V \to V$ such that $B(u,v) = \langle Bu, v\rangle$. We can fix u, then any linear functional can be written as an inner product.

Definition 3.1. Tensor product $\otimes$

Take $x, y \in V$, then $(x \otimes y)u := x\langle y, u\rangle$ for any $u \in V$. In matrix notations, $x \otimes y = xy^T$ with ij values $x_i y_j$.

Definition 3.2. Covariance operator

Recall that for $\xi, \eta \in \mathbb{R}, \mathrm{Cov}(\xi, \eta) = \mathbb{E}[(\xi - \mathbb{E}[\xi])(\eta - \mathbb{E}[\eta])]$. The map from $u, v \in V$ to $\mathrm{Cov}(\langle X, u\rangle, \langle X, v\rangle)$ is a bilinear form since linear to u and v. So there exists a linear operator $\Sigma : V \to V$ such that $\langle \Sigma u, v\rangle = \mathrm{Cov}(\langle X, u\rangle, \langle X, v\rangle)$. We call Σ be the covariance operator of X and it satisfies the following properties:

- $\Sigma u = \mathbb{E}[\langle X - \mathbb{E}[X], u\rangle (X - \mathbb{E}[X])], u \in V$
- $\Sigma = \mathbb{E}(X - \mathbb{E}[X]) \otimes (X - \mathbb{E}[X])$, or $\mathrm{Cov}(X_i, X_j)$ is the covariance matrix.

Proposition 3.2. Properties of covariance operators

1. $\Sigma = \Sigma^*$, self-adjoint.

2. Σ is positive semi-definite becasue for any $u \in V$, $\langle \Sigma u, u\rangle = \mathbb{V}\mathrm{ar}(\langle X, u\rangle) \geq 0$.

3. Any self-adjoint, positive semi-definite operator $\Sigma : V \to V$ is a covariance operator of a normal random vector.

The linear space of positive semi-definite operators (we can add or multiply the covariance operators) forms a cone, a convex set S, which means $x \in S \implies cx \in S, c \geq 0$. Or it includes non-negative multiples of vectors, which means that if we multiply the covariance operator with a non-negative number, it's still a covariance operator.

4. Let X be a random vector with covariance operator Σ_X and $T : V \to V_1$ which is a linear operator, $\Sigma_{TX} = T\Sigma_X T^*$.

Proof

For any $u, v \in V_1$, $\langle \Sigma_{TX} u, v\rangle = \mathrm{Cov}(\langle TX, u\rangle, \langle TX, v\rangle) = \mathrm{Cov}(\langle X, T^*u\rangle, \langle X, T^*v\rangle) = \langle \Sigma_X T^*u, T^*v\rangle = \langle T\Sigma_X T^*u, v\rangle \implies \Sigma_{TX} = T\Sigma_X T^*$. And the operator is uniquely defined. $\square$

Definition 3.3. Cross-covariannce operator

Suppose X is a random variable with values in the inner prodcut space V_1 and Y is a random variable with values in V_2, then define operator Σ_{XY} using the following relationship.

1. $\langle \Sigma_{XY} u, v\rangle = \mathrm{Cov}(\langle X, u\rangle, \langle Y, v\rangle)$ where $u \in V_1, v \in V_2$.

2. $\Sigma_{XY} : V_1 \to V_2$.

Proposition 3.3. Some properties for Cross-covariannce

1. $\Sigma_{YX} = \Sigma_{XY}^*$
2. $\Sigma_{XX} = \Sigma_X$, the covariacne operator X
3. If X is a random variable in V, $T_1 : V \to V_1$, $T_2 : V \to V_2$, both linear operators, then $\Sigma_{T_1 X, T_2 X} = T_2 \Sigma_{XX} T_1^*$.

Proof

$$\langle \Sigma_{T_1 X, T_2 X} u, v\rangle = \mathrm{Cov}(\langle T_1 X, u\rangle, \langle T_2 X, v\rangle) = \mathrm{Cov}(\langle X, T_1^* u\rangle, \langle X, T_2^* v\rangle) = \langle \Sigma_{XX} T_1^* u, T_2^* v\rangle = \langle T_2 \Sigma_{XX} T_1^* u, v\rangle$$

Therefore, $\Sigma_{T_1 X, T_2 X} = T_2 \Sigma_{XX} T_1^* = T_1 \Sigma_{XX} T_2^*$ since $\Sigma_{T_1 X, T_2 X}$ is self-adjoint. □

Definition 3.4. Uncorrelated

If $X \in V_1$ and $Y \in V_2$ are uncorrelated $\Leftrightarrow \forall u \in V_1, v \in V_2, \langle X, u\rangle$ and $\langle Y, v\rangle$ are uncorrelated, or $\langle \Sigma_{XY} u, v\rangle = 0, \forall u \in V_1, v \in V_2$. This is equivalent to say $\Sigma_{XY} = 0$. $T_1 X$ and $T_2 X$ are uncorrelated if and only if $\Sigma_{T_1 X, T_2 X} = T_2 \Sigma_{XX} T_1^* = 0$.

Theorem 3.1. *Suppose X is a random variable in V and Σ is a covariance opeartor of X, since Σ is self-adjoint and positively semidefinite, by spectral theorem, $\Sigma = \sum_{\lambda \in \sigma(\Sigma)} \lambda P_\lambda$. Moreover, P_λ are mutually orthogonal and is a resolution of identity.*

$I = \sum_{\lambda \in \sigma(\Sigma)} P_\lambda$, apply this to X, get $X = \sum_{\lambda \in \sigma(\Sigma)} P_\lambda X$. If we take $\lambda, \lambda' \in \sigma(\Sigma), \Sigma_{P_\lambda X, P_{\lambda'} X} = P_{\lambda'} \Sigma P_\lambda^* = P_{\lambda'} \Sigma P_\lambda = P_{\lambda'} (\sum_{\mu \in \sigma(\Sigma)} \mu P_\mu) P_\lambda$.

If $\lambda' \neq \lambda, \Sigma_{P_\lambda X, P_{\lambda'} X} = 0$.

If $\lambda' = \lambda, \Sigma_{P_\lambda X, P_{\lambda'} X} = \lambda P_\lambda$.

Corollary 3.2. *$P_\lambda X, \lambda \in \sigma(\Sigma)$ are mutually uncorrelated.*

Consider $u \in L_\lambda$, $||u|| = 1$ where $L_\lambda = Im(P_\lambda)$. Then we have $\mathbb{V}\mathrm{ar}(\langle P_\lambda X, u\rangle) = \langle \Sigma_{P_\lambda X} u, u\rangle = \langle \lambda P_\lambda u, u\rangle = \langle \lambda P_\lambda u, P_\lambda u\rangle = \lambda ||P_\lambda u||^2 = \lambda ||u||^2 = \lambda$ since P_λ projects to the eigen space.

Theorem 3.3. *Principal Component Analysis*

Let X be a random variable in V with covariance operator Σ with dimension d, then by spectral theorem, $\Sigma = \sum_{j=1}^d \lambda_j P_j$ where P_j is projection on $\mathrm{linear\ span}(l_j)$ *are orthonormal eigenvectors of Σ, $\Sigma e_j = \lambda_j e_j$ where $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_d$.*

Then we can write $X = \sum_{j=1}^{d} X_j e_j, X_j = \langle X, e_j \rangle$ *since* e_j *forms a basis of the linear space* V*.*

In addition, $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(\langle X, e_i \rangle, \langle X, e_j \rangle) = \langle \Sigma e_i, e_j \rangle = \langle \lambda_i e_i, e_j \rangle = \lambda_i \langle e_i, e_j \rangle = \lambda_i \delta_{ij}$*.*

Consequently, $X_1, \cdots, X_n$ *are uncorrelated random variables with* $\mathbb{V}\mathrm{ar}(X_j) = \lambda_j$*.*

Definition 3.5. Normal random variables in inner product spaces

Suppose X be a random variable in an inner product space V,we say X is normal or Gaussian (often in infinite dimension spaces). It means that for any $u \in V$, $\langle X, u \rangle$ is a normal random variable.

Definition 3.6. Characteristic function

If ξ is random variable in $\mathbb{R}$, the characteristic function $\phi_\xi(t) = \mathbb{E}[e^{it\xi}]$, it's well defined for $t \in \mathbb{R}$. If $\xi \sim N(\mu, \sigma^2)$, then $\phi_\xi(t) = \exp(i\mu t - \frac{\sigma^2 t^2}{2})$. If X is normal in V, then $\mathbb{E}[\langle X, u \rangle]^2$ is finite, or there exists $\mathbb{E}[X] = a$ and Σ_X. Moreover, $\mathbb{E}[\langle X, u \rangle] = \langle a, u \rangle, \mathbb{V}\mathrm{ar}(\langle X, u \rangle) = \langle \Sigma_X u, u \rangle$. It follows that the characteristic function of $\langle X, u \rangle$ is $\mathbb{E}[e^{it\langle X, u \rangle}] = \exp(i\langle a, u \rangle t - \frac{1}{2}\langle \Sigma_X u, u \rangle t^2)$.

The characteristic function is unique for each distribution, or $\phi_{X_1}(u) = \phi_{X_2}(u), u \in V \implies X_1 \overset{d}{=} X_2$.

If X is normal with mean a and covariance Σ. We have $\phi_X(u) = \mathbb{E}e^{i\langle X, u \rangle} = \exp(i\langle a, u \rangle - \frac{1}{2}\langle \Sigma u, u \rangle)$. It follows that the distribution of normal vector X is completely characterized by its mean a and covariance operator Σ.

Proposition 3.4. Suppose $X \sim N(a, \Sigma)$ in V. Let $T : V \to V_1$ be a linear operator. Then TX is normal with mean Ta and covariance $T\Sigma T^*$.

Proof

Enough to show that for any $u \in V_1$, $\langle TX, u \rangle_{V_1 \times V_1} = \langle X, T^* u \rangle_{V \times V}$ is a normal random variable. □

Theorem 3.4. *Assume* V_1, V_2 *are two inner product spaces, define the new space* V *as* $V = V_1 \oplus V_2 = \{(x_1, x_2), x_1 \in V_1, x_2 \in V_2\}$ *and* $(x_1, x_2) + (y_1, y_2) = (x_1 + x_2, y_1 + y_2), c(x_1, x_2) = (cx_1, cx_2), \langle (x_1, x_2), (y_1, y_2) \rangle = \langle x_1, y_1 \rangle + \langle x_2, y_2 \rangle$ *for some operations. Suppose* X_1 *is a random variable in* V_1 *and* X_2 *is a random variable in* V_2*, and let* $X = (X_1, X_2) \in V$*. Note that* X_1, X_2 *are linear transformations of* X*, so they are normal.*

But X_1, X_2 *are both normal does not imply that* X *is normal, see ISyE 7405 HW1 Q5 for a counter example.*

Suppose X *is normal in* V*, then the following 2 statements are equivalent.*

(1) X_1 *and* X_2 *are uncorrelated.*
(2) X_1 *and* X_2 *are independent.*

Proof

Let $X_1 \sim N(a_1,\Sigma_1), X_2 \sim N(a_2,\Sigma_2), X \sim N(a,\Sigma)$. $a=(a_1,a_2)$.

Define $\langle \Sigma u, v\rangle = \mathrm{Cov}(\langle X,u\rangle,\langle X,v\rangle) = \mathrm{Cov}(\langle X_1,u_1\rangle + \langle X_2,u_2\rangle, \langle X_1,v_1\rangle + \langle X_2,v_2\rangle)$, which is equal to $\mathrm{Cov}(\langle X_1,u_1\rangle,\langle X_1,v_1\rangle) + \mathrm{Cov}(\langle X_1,u_1\rangle,\langle X_2,v_2\rangle) + \mathrm{Cov}(\langle X_2,u_2\rangle,\langle X_1,v_1\rangle) + \mathrm{Cov}(\langle X_2,u_2\rangle,\langle X_2,v_2\rangle) = \langle \Sigma_{X_1X_1}u_1,v_1\rangle + \langle \Sigma_{X_1X_2}u_1,v_2\rangle + \langle \Sigma_{X_2X_1}u_2,v_1\rangle + \langle \Sigma_{X_2X_2}u_2,v_2\rangle$.

Or $\langle \Sigma u, v\rangle = \langle \Sigma_{X_1X_1}u_1,v_1\rangle + \langle \Sigma_{X_1X_2}u_1,v_2\rangle + \langle \Sigma_{X_2X_1}u_2,v_1\rangle + \langle \Sigma_{X_2X_2}u_2,v_2\rangle$.

We can think Σ to be the following operator

$$\langle \Sigma u, v\rangle = \langle \begin{bmatrix} \Sigma_{X_1X_1} & \Sigma_{X_1X_2} \\ \Sigma_{X_2X_1} & \Sigma_{X_2X_2} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \rangle$$

X_1 and X_2 are uncorrelated, so $\Sigma_{X_1X_2} = \Sigma_{X_2X_1} = 0$. It follows that $\langle \Sigma u, v\rangle = \langle \Sigma_1 u_1, v_1\rangle + \langle \Sigma_2 u_2, v_2\rangle$.

And the characteristic function of X is $\phi_X(u) = \exp(i\langle a,u\rangle - \frac{1}{2}\langle \Sigma u,u\rangle) = \exp(i\langle a_1,u_1\rangle - \frac{1}{2}\langle \Sigma_1 u_1,u_1\rangle + i\langle a_2,u_2\rangle - \frac{1}{2}\langle \Sigma_2 u_2,u_2\rangle) = \exp(i\langle a_1,u_1\rangle - \frac{1}{2}\langle \Sigma_1 u_1,u_1\rangle)\cdot \exp(i\langle a_2,u_2\rangle - \frac{1}{2}\langle \Sigma_2 u_2,u_2\rangle) = \phi_{X_1}(u_1)\cdot\phi_{X_2}(u_2)$

Let Y_1, Y_2 be independent random variables, and $Y_1 \sim N(a_1,\Sigma_1), Y_2 \sim N(a_2,\Sigma_2)$, and $Y=(Y_1,Y_2)$. Then $\phi_Y(u) = \mathbb{E}e^{i\langle Y,u\rangle} = \mathbb{E}e^{i(\langle Y_1,u_1\rangle + \langle Y_2,u_2\rangle)} = \mathbb{E}(e^{i\langle Y_1,u_1\rangle}e^{i\langle Y_2,u_2\rangle}) = \mathbb{E}e^{i\langle Y_1,u_1\rangle}\cdot\mathbb{E}e^{i\langle Y_2,u_2\rangle} = \phi_{Y_1}(u_1)\phi_{Y_2}(u_2) = \phi_{X_1}(u_1)\phi_{X_2}(u_2) = \phi_X(u) \implies X \overset{d}{=} Y$. Therefore, X_1 and X_2 are independent.

$\square$

Corollary 3.5. *Let $X \sim N(a,\Sigma)$, the spectral representation is $\Sigma = \sum_{j=1}^d \lambda_j P_j$ where $P_j = e_j \otimes e_j$ ($e_1 \cdots e_d$ are orthonormal vectors) are orthogonal projection on* linear span(e_j) *where $\lambda_1 \geq \cdots \geq \lambda_d$ and e_j are orthonormal vectors, $\Sigma e_j = \lambda_j e_j$.*

We can write $X = \sum_{\lambda\in\sigma(\Sigma)} P_\lambda X$. $P_\lambda X$ and $P_{\lambda'} X$ are uncorrelated, since normal, they are independent.

Moreover, $P_\lambda X \sim N(P_\lambda a, \lambda P_\lambda)$. Note that $P_\lambda X \in L_\lambda = Im(P_\lambda)$, in $Im(P_\lambda), P_\lambda X \sim N(P_\lambda a, \lambda I_{L_\lambda})$.

$X = \sum_{j=1}^n X_j e_j, X_j = \langle X, e_j\rangle$ and different X_i are uncorrelated (can be checked by definition of covariance operators, also independent) and $\mathbb{V}\mathrm{ar}(X_j) = \mathbb{V}\mathrm{ar}(\langle X,e_j\rangle) = \langle \Sigma e_j, e_j\rangle = \langle \lambda_j e_j, e_j\rangle = \lambda_j = \sigma_j^2$.

Let $a = \sum_{j=1}^n a_j e_j$, then $\mathbb{E}\langle X, e_j\rangle = a_j$ and $X_j \sim N(a_j, \sigma_j^2)$.

$P_{X_1\cdots X_d}(x_1\cdots x_d) = P_{X_1}(x_1)\cdots P_{X_d}(x_d)$ by independence of components, and we need variance of each component to be positive.

Therefore, $P_{X_1\cdots X_d}(x_1\cdots x_d) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x_j-a_j)^2}{2\sigma_j^2}} = \frac{1}{(2\pi)^{d/2}\sigma_1\cdots\sigma_d} e^{-\frac{1}{2}\sum_{j=1}^d \frac{(x_j-a_j)^2}{\sigma_j^2}}$

$$= \frac{1}{(2\pi)^{d/2}\lambda_1\cdots\lambda_d} e^{-\frac{1}{2}\sum_{j=1}^d \frac{(x_j-a_j)^2}{\lambda_j^2}} = \frac{1}{(2\pi)^{\frac{d}{2}}\det(\Sigma)} e^{-\frac{1}{2}\langle \Sigma^{-1}(x-a),(x-a)\rangle}$$

Since $\sum_{j=1}^d \frac{(x_j-a_j)^2}{\lambda_j} = \langle \Sigma^{-1}(x-a),(x-a)\rangle$ ($\langle \Sigma u, u\rangle = \sum_{j=1}^d \lambda_j u_j^2$ and $\langle \Sigma^{-1}u,u\rangle = \sum_{j=1}^d \lambda_j^{-1} u_j^2$) since change of basis does not change the inner product.

Definition 3.7. Chi-square χ^2 distribution

Let $Z_1, \cdots, Z_n$ i.i.d $N(0,1)$, then $Z_1^2 + \cdots Z_d^2$ follows a Chi-square distribution with degree of freedom d, or χ_d^2.

Take any $\mu \geq 0$, and write $(Z_1+\mu)^2 + Z_2^2 + \cdots Z_d^2) \sim \chi_{d,\mu}^2$ (non-central chi-square distribution).

Let $X \sim N(\mu,1), X = \mu + Z$ where Z is standard normal,

$$\mathbb{E}e^{tX^2} = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{tx^2} e^{-\frac{(x-\mu)^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{tx^2} e^{-\frac{x^2}{x}} e^{x\mu} e^{-\frac{\mu^2}{2}} dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2}} \int_{\mathbb{R}} e^{-\frac{1}{2}(1-2t)x^2} e^{x\mu} dx$$

Which is the MGF of $N(0, \sigma^2 = \frac{1}{1-2t})$ at some value, $\mathbb{E}_Z e^{\mu\sigma Z} = e^{\frac{\mu^2\sigma^2}{2}}$ where Z is standard normal since $M_X(t) = \exp(\mu t + \frac{\sigma^2 t^2}{2})$ if $X \sim N(\mu, \sigma^2)$.

Therefore,

$$\mathbb{E}e^{tX^2} = e^{-\frac{\mu^2}{2}} \underbrace{\sigma \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(1-2t)x^2} e^{x\mu} dx}_{\exp(\frac{\mu^2\sigma^2}{2})} = e^{-\frac{\mu^2}{2}} e^{\frac{\mu^2}{2(1-2t)}} \frac{1}{\sqrt{1-2t}}$$

MGF for $\chi_d^2 = \mathbb{E}e^{t(Z_1^2+\cdots+Z_d^2)} = \prod_{i=1}^d \mathbb{E}e^{tZ_d^2} = \frac{1}{(1-2t)^{\frac{d}{2}}}$.

MGF for $\chi_{d,\mu}^2 = \mathbb{E}e^{t((Z_1+\mu)^2 \cdots Z_d^2)} = \mathbb{E}e^{t(Z_1+\mu)^2} \prod_{i=2}^d \mathbb{E}e^{tZ_d^2} = e^{-\frac{\mu^2}{2}} e^{\frac{\mu^2}{2(1-2t)}} \frac{1}{(1-2t)^{\frac{d}{2}}} = e^{\frac{\mu^2 t}{1-2t}} (1-2t)^{-\frac{d}{2}}$.

The Taylor expansion for $e^{\frac{\mu^2}{2(1-2t)}}$ is $\sum_{k=0}^\infty \frac{(\frac{\mu^2}{2})^k}{k!} \frac{1}{(1-2t)^k}$.

So we have $\mathbb{E}e^{t((Z_1+\mu)^2+\cdots+Z_d^2)} = \sum_{k=0}^\infty e^{-\frac{\mu^2}{2}} \frac{(\frac{\mu^2}{2})^k}{k!} \frac{1}{(1-2t)^{\frac{2k+d}{2}}} = \sum_{k=0}^\infty e^{-\frac{\mu^2}{2}} \frac{(\frac{\mu^2}{2})^k}{k!} \mathbb{E}e^{t(Z_1^2+\cdots+Z_{2k+d}^2)}$.

The first term is PGF of Poisson distribution with parameter $\frac{\mu^2}{2}$, and the second term is the MGF of χ_{2k+d}^2. We also have $F_{\chi_{d,\mu}^2} = \sum_{k=0}^\infty e^{-\frac{\mu^2}{2}} \frac{(\frac{\mu^2}{2})^k}{k!} F_{\chi_{d+2k}^2}$ since there is a one-to-one relation between CDF and MGF.

We start with non-central distribution, but we can view it as a Poisson mixture of Chi-square distribution.

Definition 3.8. $\mathscr{F}$ distribution

Consider $S_1 \sim \chi_{d_1,\mu}^2, S_2 \sim \chi_{d_2}^2$, and S_1 and S_2 are independent, then $\frac{S_1}{S_2} \sim \mathscr{F}_{d_1,d_2,\mu}$.

Proposition 3.5. Suppose $Z \sim N(0, I_d)$ in V with dimension d, then $||Z||^2 = Z_1^2 + \cdots + Z_d^2 \sim \chi_d^2$, where Z_i are i.i.d. standard normal variable.

Suppose $A : V \to V$, then if A is self-adjoint, we can create quadratic form of $A : \langle AZ, Z \rangle$, if $Z \sim N(0, I)$.

We can write spectral decomposition of $\langle AZ,Z\rangle$, suppose $A=\sum_{k=1}^{d}\lambda_k e_k\otimes e_k$, then $\langle AZ,Z\rangle=\sum_{j=1}\lambda_j Z_j^2$ where $Z_j=\langle Z,e_j\rangle$ (since change of basis does not change the inner product).

The MGF is $\mathbb{E}e^{t\langle AZ,Z\rangle}=\mathbb{E}e^{t(\sum_{j=1}^{d}\lambda_j Z_j^2)}=\prod_{j=1}^{d}\mathbb{E}e^{t\lambda_j Z_j^2}=\prod_{j=1}^{d}\frac{1}{1-2\lambda_j t}=\sqrt{\frac{1}{\prod_{j=1}^{d}(1-2\lambda_j t)}}=\sqrt{\frac{1}{\det(I-2tA)}}$ if $2\lambda_j t<1$ for all j since $1-2\lambda_j t$ are eigenvalues of $I-2tA$.

Proposition 3.6. Suppose $Z\sim N(0,I_d)$ in V with dimension d, and if $A:V\to V$ is self-adjoint, then $\langle AZ,Z\rangle\sim\chi_k^2\Leftrightarrow A=P_L, L\subset V, \dim(L)=k, k\leq d$.

Proof

Assume that A has eigenvalues λ_k in decreasing order, with corresponding eigenvectors e_k, then $\mathbb{E}e^{t\langle AZ,Z\rangle}=\prod_{i=1}^{d}\frac{1}{\sqrt{1-2\lambda_i t}}=\frac{1}{(1-2t)^{k/2}}(MGF\ of\ \chi_k^2)$.

$\implies \prod_{i=1}^{d}(1-2\lambda_i t)=(1-2t)^k$

The are polynomials, so they have the same roots, so $\lambda_j=1, j\leq k$ and $\lambda_j=0, j>k$

$\implies A=\sum_{j=1}^{d}\lambda_j e_j\otimes e_j=\sum_{j=1}^{k}e_j\otimes e_j=P_L$ where $L=$ linear span$(e_1,\cdots,e_k)$ by proposition 2.6.

$\square$

Proposition 3.7. If $X\sim N(a,I)$ in V with $\dim(V)=d$, then $||X||^2\sim\chi^2_{d,||a||}$.

Proof

Choose $v=\frac{a}{||a||}$ then $e_1=v, e_2,\cdots,e_d$ are the orthonormal bases.

$X=a+Z=||a||e_1+\langle Z,e_1\rangle e_1+\cdots+\langle Z,e_d\rangle e_d$ where $Z_j=\langle Z,e_j\rangle\sim N(0,1)$, where this follows from corollary 3.5. Note that I has eigenvalues 1 and $\mu_Z=a=0$.

Then $||X||^2=||(a+Z_1)e_1+Z_2e_2+\cdots Z_de_d||^2=(||a||+Z_1)^2+Z_2^2+\cdots+Z_d^2\sim\chi^2_{d,||a||}$. $\square$

Corollary 3.6. *Let $X\sim N(a,I)$ in V, $d=dim(V)$, $L\subset V$ is a subset of V.* $||P_LX||^2\sim\chi^2_{\dim(L),||P_La||}$.

Proof

$P_L X \sim N(P_L a, P_L)$. In space L, $P_L X \sim N(P_L a, I_L)$. $\square$

Corollary 3.7. *If $X \sim N(a, \sigma^2 I)$ in V, then $\frac{X}{\sigma} \sim N(\frac{a}{\sigma}, I)$, and $||P_L X||^2 = \sigma^2 ||P_L \frac{X}{\sigma}||^2 \sim \sigma^2 \chi^2_{\dim(L), \frac{||P_L a||}{\sigma}}$.*

If $Z \sim N(0, I)$ in V with dimension d. Consider arbitrary A, then $\langle AZ, Z\rangle = \langle Z, A^* Z\rangle = \langle A^* Z, Z\rangle \implies \langle AZ, Z\rangle = \frac{1}{2}(\langle AZ, Z\rangle + \langle A^* Z, Z\rangle) = \langle \underbrace{\frac{A+A^*}{2}}_{\text{self-adjoint}} Z, Z\rangle$.

So for quadratic forms, only considering self-adjoint operators is enough.

Use spectral decomposition of $A = \sum_{k=1}^d \lambda_k e_k \otimes e_k$.

So $\mathbb{E}\langle AZ, Z\rangle = \mathbb{E}\sum_{k=1}^d \lambda_k Z_k^2 = \sum_{k=1}^d \lambda_k \mathbb{E} Z_k^2 = \sum_{k=1}^d \lambda_k = tr(A)$. It's true for normal Z, but also for arbitrary Z with $\mathbb{E}Z = 0$ and $\Sigma = I$.

Now to get variance, $\mathbb{V}\text{ar}(\langle AZ, Z\rangle) = \mathbb{V}\text{ar}(\sum_{k=1}^d \lambda_k Z_k^2) = \sum_{k=1}^d \mathbb{V}\text{ar}(\lambda_k Z_k^2) = \sum_{k=1}^d \lambda_k^2 \mathbb{V}\text{ar}(Z_k^2) = \sum_{k=1}^d \lambda_k^2 2 = 2tr(A^2) = 2tr(AA) = 2||A||_2^2$ (which is called the Hilbert-Schmidt norm) since the fourth central moment of a normal distribution is $3\sigma^4$.

If $X \sim N(0, \Sigma)$, then consider $X = \Sigma^{\frac{1}{2}} Z$.

Then $\mathbb{E}\langle AX, X\rangle = \mathbb{E}\langle A\Sigma^{\frac{1}{2}} Z, \Sigma^{\frac{1}{2}} Z\rangle = \mathbb{E}\langle \Sigma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}} Z, Z\rangle = tr(\Sigma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}}) = tr(A\Sigma) = tr(\Sigma A)$.

And $\mathbb{V}\text{ar}(\langle AX, X\rangle) = \mathbb{V}\text{ar}(\langle \Sigma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}} Z, Z\rangle) = 2tr(\Sigma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}} A) = 2tr(A\Sigma A\Sigma) = 2tr(\Sigma A \Sigma A)$.

If $X \sim N(\mu, \Sigma)$, then consider $X = \Sigma^{\frac{1}{2}} Z + \mu$. It can be shown that $\mathbb{E}\langle AX, X\rangle = tr(A\Sigma) + \mu^* A \mu$ and $\mathbb{V}\text{ar}(\langle AX, X\rangle) = 2tr(\Sigma A \Sigma A) + 4\mu^* A \Sigma A \mu$.

Definition 3.9. Weakly (strongly) spherical or isotropic vector

If X is a random vector V, $\mathbb{E}X = a, \Sigma_X = \sigma^2 I$, then X is called a weakly spherical or isotropic, and strongly spherical basically means normal variable.

If $L_1 \cdots L_k$ are subspaces and $L_i \perp L_j, i \neq j$, then $P_{L_1} X, \cdots, P_{L_k} X$ are uncorrelated random variable. Also for any j, $P_{L_j} X$ is weakly spherical in L_j.

In addition, $\mathbb{E}||P_j X||^2 = \sigma^2 \dim(L_j) + ||P_{L_j} a||^2$, the proof is quite similar to proposition 3.7. Just use bases $e_1 = \frac{P_{L_j} a}{||P_{L_j} a||}$ and $e_2, \cdots, e_{\dim L_j}$.

Another fact is that if X is normal with mean a and variance $\sigma^2 I$, (strongly spherical), $L_1 \cdots L_k \subset V$, $L_i \perp L_j$ when $i \neq j$. $P_{L_1}(X), \cdots, P_{L_k} X$ will be uncorrelated hence independent. Each of $P_{L_j}(X) \sim N(P_{L_j} a, P_{L_j})$.

Therefore, $||P_{L_j} X||^2, j = 1, \cdots, k$ are also independent and $||P_{L_j} X||^2 \sim \sigma^2 \chi_{\dim(L_j), \frac{||P_{L_j} a||}{\sigma}}$.

Theorem 3.8. *Cochran's theorem*

Suppose $X \sim N(a, \sigma^2 I)$ in V, and $A, A_1, \cdots, A_k : V \to V$ are self-adjoint operators, and $A = A_1 + \cdots + A_k$.

If A is an orthogonal projection, then the following statements are equivalent:

(i) A_i is orthogonal projection for any i.

(ii) $A_i A_j = 0$ if $i \neq j$.

(iii) $Im(A_i) \perp Im(A_j)$, $i \neq j$.

(iv) $rank(A)$ (the dimension of $Im(A)$) is equal to $rank(A_1) + \cdots + rank(A_k)$.

Moreover, if any of the conditions hold, then the quadratic forms $\langle A_i X, X\rangle \sim \sigma^2 \chi^2_{rank(A_i), \frac{\langle A_i a, a\rangle^{1/2}}{\sigma}}$. Moreover, these quadratic forms are independent. Note that $\langle A_i X, X\rangle = ||A_i X||^2$ since A_i is a projection. Also, $||P_L(a)|| = ||Aa|| = \langle Aa, Aa\rangle^{\frac{1}{2}} = \langle A^2 a, a\rangle^{\frac{1}{2}} = \langle Aa, a\rangle^{\frac{1}{2}}$.

*Remark: $A = P_L$ is orthogonal projection it's **equivalent** to say $\langle AX, X\rangle \sim \sigma^2 \chi^2_{rank(A), \frac{\langle Aa, a\rangle^{1/2}}{\sigma}}$. So we can change the condition that A is orthogonal projection to $\langle AX, X\rangle \sim \sigma^2 \chi^2_{rank(A), \frac{\langle Aa, a\rangle^{1/2}}{\sigma}}$.*

We've proved it's true for mean of 0, but it's still true for any mean.

Example 3.1. Assume $X_1, \cdots X_n$ are iid $N(\mu, \sigma^2)$, to test the hypothesis $H_0 : \mu = 0$ and $H_1 : \mu \neq 0$

The student t statistics is defined as $T = \sqrt{n}\frac{\bar{X}}{S} \sim t_{n-1}$ under H_0 where $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$.

1. $\bar{X}$ and S are independent random variables.

2. $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

3. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

4. $T \overset{d}{=} \frac{\sigma Z}{\sigma\sqrt{\frac{\chi^2_{n-1}}{n-1}}} = \frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$ follows t_{n-1} (numerator and denominator are independent).

5. Under H_1, T follows non-central t distribution, can be used to calculate type 2 error.

If $|T| \geq t_{\frac{\alpha}{2}}$, we reject H_0, otherwise not reject.

Derivation from Cochran theorem:

We have $\sum_{i=1}^n X_i^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n\bar{X}^2$.

Write $X = (X_1, \cdots, X_n)$, then $X \sim N(a, \sigma^2 I_n)$ where $a = (\mu, \cdots, \mu) \in \mathbb{R}^n$.

Define 3 quadratic terms, $Q(X) = \sum_{j=1}^n X_j^2 = ||X||^2 = \langle I_n X, X\rangle$ and $Q_1(X) = \sum_{i=1}^n (X_i - \bar{X})^2 = \langle A_1 X, X\rangle$. The A_1 always exists (we can always write quadratic terms in self-adjoint operators) and $Q_2(X) = n\bar{X}^2 = \langle A_2 X, X\rangle$ where A_1 and A_2 are self-adjoint and positive semi-definite, plus, $Q(X) = Q_1(X) + Q_2(X) \implies I_n = A_1 + A_2$ since we can get the bilinear form (one-to-one to operators) from quadratic form.

$Im(A_1) = Ker(A_1)^{\perp}$ (since self-adjoint), $Q_1(X) = 0 \Leftrightarrow \langle A_1X, X\rangle = 0 \Leftrightarrow \langle A_1^{\frac{1}{2}}X, A_1^{\frac{1}{2}}X\rangle \Leftrightarrow ||A_1^{\frac{1}{2}}X||^2 = 0 \Leftrightarrow A_1^{\frac{1}{2}}X = 0 \implies A_1X = 0 \Leftrightarrow X \in Ker(A_1) \Leftrightarrow Q_1(X) = 0$ so $Q_1(X) = 0 \Leftrightarrow X \in Ker(A_1)$ or $Q_1(X) = 0 \Leftrightarrow X_j = \bar{X}, j = 1, \cdots, n$ or we only have $n-1$ independent equations. This implies that $\dim(Ker(A_1)) = 1 \implies \dim(Im(A_1)) = n-1 \implies rank(A_1) = n-1$.

$Q_2(X) = 0 \Leftrightarrow \langle A_2X, X\rangle = 0 \Leftrightarrow X \in Ker(A_2)$.

$Q_2(X) = 0 \Leftrightarrow \bar{X} = 0$, which is a hyperplane $X_1 + \cdots + X_n = 0$ with $\dim(Ker(A_2)) = n-1$ and $rank(A_2) = \dim(Im(A_2)) = 1$.

So $rank(A) = rank(A_1) + rank(A_2)$.

It follows that A_1 and A_2 are orthogonal projections that $\langle A_1X, X\rangle$ and $\langle A_2X, X\rangle$ are independent where $\langle A_1X, X\rangle \sim \sigma^2\chi^2_{rank(A_1)=n-1, \frac{\langle A_1\mu,\mu\rangle^{1/2}}{\sigma} = \frac{Q_1(\mu)}{\sigma} = 0} = \sigma^2\chi^2_{n-1}$.

In addition, $\langle A_2X, X\rangle \sim \sigma^2\chi^2_{rank(A_2)=1, \frac{\langle A_2\mu,\mu\rangle^{1/2}}{\sigma} = \frac{\sqrt{n\mu^2}}{\sigma} = \frac{\sqrt{n}|\mu|}{\sigma}}$.

So 1) $\sum_{j=1}^n (X_j - \bar{X})^2$ and $n\bar{X}^2$ are independent random variables.

2) $\sum_{j=1}^n (X_j - \bar{X})^2 \sim \sigma^2\chi^2_{n-1}$ (Pearson theorem).

3) $n\bar{X}^2 \sim \sigma^2\chi^2_{1, \frac{\sqrt{n}|\mu|}{\sigma}}$. Under H_0, $\frac{n\bar{X}^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \stackrel{d}{=} \frac{\chi_1^2}{\chi^2_{n-1}} \sim \mathscr{F}_{1,n-1}$ which reduces to the square of student-t test.

Example 3.2. $X_{ij}, i = 1, \cdots, m, j = 1, 2 \cdots n_i$ iid $N(\mu_i, \sigma^2)$ and $n = n_1 + \cdots n_m$.

m samples from normal distribution with possibly different means and the same variance.

$H_0 : \mu_1 = \cdots \mu_m$, H_1: otherwise, exists two or more different means.

Denote $\bar{X}_i$ be the sample mean for sample i, $\frac{X_{i1} + \cdots + X_{in_i}}{n_i}$ and $S_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$.

$\bar{X} = \frac{\sum_{i,j} X_{ij}}{n}$, $S^2 = \frac{\sum_{i=1}^m n_i S_i^2}{\sum_{i=1}^m n_i}$, no need to normalize S for now.

H_0 is equivalent to the equality $\sum_{i=1}^m n_i(\mu_i - \bar{\mu})^2 = 0$. We can create an estimator for this, $\sum_{i=1}^m n_i(\bar{X}_i - \bar{X})^2$.

Identity: $\sum_{i=1}^m \sum_{j=1}^{n_i} X_{ij}^2 = \sum_{i=1}^m n_i S_i^2 + \sum_{i=1}^m n_i(\bar{X}_i - \bar{X})^2 + n\bar{X}^2$.

$X = (X_{ij}) \in \mathbb{R}^n \sim N(a, \sigma^2 I_n)$.

a is a long vector with n_1 values of $\mu_1, \cdots n_m$ values of μ_m, i.e. $a = (\underbrace{\mu_1, \cdots, \mu_1}_{n_1} \underbrace{\mu_2, \cdots, \mu_2}_{n_2}, \cdots, \underbrace{\mu_m, \cdots, \mu_m}_{n_m})'$.

$Q(X) = ||X||^2 = Q_1(X) + Q_2(X) + Q_3(X)$ where $Q_1(X) = \langle A_1X, X\rangle = \sum_{i=1}^m n_i S_i^2$, $Q_2(X) = \langle A_2X, X\rangle = \sum_{i=1}^m \sum_{i=1}^m n_i(\bar{X}_i - \bar{X})^2$, $Q_3(X) = \langle A_3X, X\rangle = n\bar{X}^2$ and A_i are self-adjoint, and positive semi-definite, $I_n = A_1 + A_2 + A_3$.

$Q_1(X) = 0 \Leftrightarrow X \in Ker(A_1) \Leftrightarrow S_j^2 = 0, i = 1, \cdots, m \Leftrightarrow X_{ij} = \bar{X}_i, i = 1, \cdots, m$, we have $n-m$ linear independent equations, the dimension of the kernel of $A_1 = m$, so $rank(A_1) = n-m$.

Similarly, $rank(A_2) = m-1, rank(A_3) = 1$. (same as the last example) So $rank(A_1) + rank(A_2) + rank(A_3) = n = rank(I_n)$.

It follows Cochran theorem that $\sum_{i=1}^m n_i S_i^2$, $\sum_{i=1}^m n_i(\bar{X}_i - \bar{X})^2$, $n\bar{X}^2$ are independent random variables.

$\sum_{i=1}^m n_i S_i^2 \sim \sigma^2 \chi^2_{n-m, \frac{\sqrt{\langle A_1 a, a\rangle}}{\sigma}=0}$.

$\sum_{i=1}^m n_i(\bar{X}_i - \bar{X})^2 \sim \sigma^2 \chi^2_{m-1, \frac{\sqrt{\langle A_2 a, a\rangle}}{\sigma}=\frac{\sqrt{\sum_{i=1}^m n_i(\mu_i-\bar{\mu})^2}}{\sigma}}$ where $\bar{\mu} = \frac{\sum_{i=1}^m n_i\mu_i}{\sum_{i=1}^m n_i}$ is the weighted average of μ_i.

The test statistics is based on

$$\frac{\sum_{i=1}^m n_i S_i^2}{\sum_{i=1}^m n_i(\bar{X}_i - \bar{X})^2} \overset{d}{=} \frac{\chi^2_{m-1, \frac{\sqrt{\sum_{i=1}^m n_i(\mu_i-\bar{\mu})^2}}{\sigma}}}{\chi^2_{n-m}} \sim \mathscr{F}_{m-1, n-m, \frac{\sqrt{\sum_{i=1}^m n_i(\mu_i-\bar{\mu})^2}}{\sigma}}$$

So under H_0, $\frac{\sum_{i=1}^m n_i S_i^2}{\sum_{i=1}^m n_i(\bar{X}_i - \bar{X})^2} \sim \mathscr{F}_{m-1,n-m}$.

4 Linear Models

Our model form is $\mathbf{Y} = \mathbf{X}\beta^* + \xi$ with unkown noise.

Basic assumption:

$Y \in V, \beta \in W$ inner product space.

$X : W \to V$, a linear operator. $Y \in V, \xi \in V$

$\mathbb{E}\xi = 0, \Sigma_\xi = \sigma^2 I_V$.

$\hat{\beta} = \text{argmin}_{u\in W} ||Y - Xu||^2$. $\hat{\beta}$ is not unique, we can add any kernel of X to $\hat{\beta}$.

Let $\mu = X\beta$, then $Y = \mu + \xi, \mu \in Im(X) = L \subset V$, called the **random shift model**.

$\hat{\mu} = \text{argmin}_{u\in L} ||Y - u||^2 = P_L Y$. So $\hat{\mu}$ is the estimator of μ.

We can write $\hat{\beta}$ is a LS-estimator which means that $X\hat{\beta} = \hat{\mu}$. Or $X\hat{\beta} = P_L Y \Leftrightarrow Y - X\hat{\beta} \perp L \Leftrightarrow Y - X\hat{\beta} \perp Xu, u \in W \Leftrightarrow \langle Y - X\hat{\beta}, Xu\rangle = 0 \Leftrightarrow \langle X\hat{\beta}, Xu\rangle = \langle Y, Xu\rangle, u \in W \Leftrightarrow \langle X^*X\hat{\beta}, u\rangle = \langle X^*Y, u\rangle, u \in W \Leftrightarrow X^*X\hat{\beta} = X^*Y$, which is called the **normal equation**.

So $\hat{\beta} \in X^+Y + Ker(X)$, and X^+ is the Moore-Penrose pseudoinverse of X. If X^*X is nonsingular, then we will ahve $\hat{\beta} = (X^*X)^{-1}X^*Y$ as the unique solution of the LS problem.

Esimation of liner function of μ: Let $f(\mu) = \langle \mu, c\rangle, \mu \in L \subset V$, c could be any vector in L (WLOG, otherwise $c \to P_L c$, or add anything orthogonal to L will get 0, $\langle \mu, c\rangle = \langle \mu, P_L c\rangle$, so only defining on L instead of V is OK).

Our goal is to estimate $f(\mu)$ based on Y.

Plug-in estimator: $\langle \hat{\mu}, c\rangle = \langle P_L Y, c\rangle = \langle Y, c\rangle$ for $c \in L$. Can we do any better?

Theorem 4.1. *Gauss-Markov theorem*

*Suppose $\langle Y,d\rangle$ for some $d\in V$ is a linear (of Y), **unbiased estimator** of linear functional $\langle\mu,c\rangle,\mu\in L$, then the claim is that $\mathbb{V}\mathrm{ar}(\langle Y,d\rangle)\geq\mathbb{V}\mathrm{ar}(\langle\hat{\mu},c\rangle),\mu\in L$. Plus, $\langle\hat{\mu},c\rangle$ is the unique linear unbiased estimator with the smallest possible variance. $\langle\hat{\mu},c\rangle$ is **BLUE** (the best linear unbiased estimator).*

Proof

$\langle Y,d\rangle$ is unbiased so $\mathbb{E}\langle Y,d\rangle=\langle\mu,c\rangle,\mu\in L$

$\mathbb{E}\langle Y,d\rangle=\langle\mathbb{E}X,d\rangle=\langle\mu,d\rangle\implies\langle\mu,c\rangle=\langle\mu,d\rangle,\mu\in L\implies d-c\perp L,c\in L$ (WLOG). $\implies c=P_Ld$.

$\mathbb{V}\mathrm{ar}(\langle Y,d\rangle)=\langle\Sigma_Y d,d\rangle=\sigma^2\langle I_V d,d\rangle=\sigma^2||d||^2\geq\sigma^2||P_Ld||^2=\sigma^2||c||^2=\sigma^2\langle Ic,c\rangle=\langle\Sigma_Y c,c\rangle=\mathbb{V}\mathrm{ar}(\langle y,\underbrace{c}_{\in L}\rangle)=$

$\mathbb{V}\mathrm{ar}(\langle P_Ly,c\rangle)=\mathbb{V}\mathrm{ar}(\langle\hat{\mu},c\rangle)$ where we used the fact that projection should be shorter.

To have equality, we need $d=P_Ld\implies d=c\implies\langle Y,d\rangle=\langle\hat{\mu},c\rangle$. $\square$

A more general problem is that suppose now $C:L\to V_1$ (arbitrary space, a linear operator). Our goal is to estimate $C\mu$ for $\mu\in L$. Again, the plug-in estimator will be $C\hat{\mu}$.

Corollary 4.2. *Suppose D is a mapping from V into V_1 which is a linear operator, and DY is an unbiased estimator of $C\mu,\mu\in L$, then we will have $\Sigma_{DY}\geq\Sigma_{C\hat{\mu}}$. (matrix $A\geq B$ means $A-B$ is positive semi-definite, in other words, $\langle(A-B)u,u\rangle\geq 0$).*

Proof

Unbiased means $\mathbb{E}DY=C\mu,\mu\in L\Leftrightarrow D\mu=C\mu,\mu\in L$. Take any inner product, we get $\langle D\mu,u\rangle=\langle C\mu,u\rangle\implies\langle\mu,D^*u\rangle=\langle\mu,C^*u\rangle\implies\langle Y,D^*u\rangle$ is an unbiased estimator of $\langle\mu,C^*u,\rangle$, or $\mathbb{E}\langle Y,D^*u\rangle=\langle\mu,D^*u\rangle=\langle\mu,C^*u\rangle$.

We need the following inequality: $\Sigma_{Dy}\geq\Sigma_{C\hat{\mu}}\Leftrightarrow\langle\Sigma_{DY}u,u\rangle\geq\langle\Sigma_{C\hat{\mu}}u,u\rangle\Leftrightarrow\mathbb{V}\mathrm{ar}(\langle DY,u\rangle)\geq\mathbb{V}\mathrm{ar}(\langle C\hat{\mu},u\rangle)\Leftrightarrow\mathbb{V}\mathrm{ar}(\langle Y,D^*u\rangle)\geq\mathbb{V}\mathrm{ar}(\langle\hat{\mu},C^*u\rangle)$ for any $u\in L$.

The last inequality holds since $\langle Y,D^*u\rangle$ is an unbiased estimator of $\langle\mu,C^*u,\rangle$ (reduction of Gauss-Markov) $\square$

Proposition 4.1. Let ξ be a random vector with mean $\mathbf{0}$, $\Sigma_\xi=\Sigma$, then $\mathbb{E}||\xi||^2=tr(\Sigma_\xi)$.

Proof

$||\xi||^2=\sum_{j=1}^d\langle\xi,e_j\rangle^2\implies\mathbb{E}||\xi||^2=\sum_{j=1}^d\mathbb{E}\langle\xi,e_j\rangle^2=\sum_{j=1}^d\mathbb{V}\mathrm{ar}(\langle\xi,e_j\rangle)=\sum_{j=1}^d\langle\Sigma e_j,e_j\rangle=\sum_{j=1}^d\langle\lambda_j e_j,e_j\rangle=\sum_{j=1}^d\lambda_j$

We get $\mathbb{E}\langle\xi,e_j\rangle^2=\mathbb{V}\mathrm{ar}(\langle\xi,e_j\rangle)$ since $\mathbb{E}\xi=0$.

□

Corollary 4.3. *Let $D : V \to V_1$ a linear operator, DY is an unbiased estimator of $C\mu, \mu \in L$, then* $\mathbb{E}||DY - C\mu||^2 \geq \mathbb{E}||C\hat{\mu} - C\mu||^2, \mu \in L$

Proof

We know that $\Sigma_{DY} \geq \Sigma_{C\hat{\mu}}$ by the previous corollary.

For any $A \geq B$, $\sum_i \langle Ae_i, e_i \rangle = \sum_i \langle \lambda_i e_i, e_i \rangle = \sum_i \lambda_{Ai} \geq \sum_i \langle Be_i, e_i \rangle = \sum_i \lambda_{Bi}$.

Set $A = \Sigma_{DY}$ and $B = \Sigma_{C\hat{\mu}}$, we have $tr(\Sigma_{DY}) \geq tr(\Sigma_{C\hat{\mu}})$.

By the preceding proposition, $tr(\Sigma_{DY}) = \mathbb{E}||DY - C\mu||^2$ and $tr(\Sigma_{C\hat{\mu}}) = \mathbb{E}||C\hat{\mu} - C\mu||^2$. □

In particular, this applies to the case where $V_1 = V, C = I$. For any linear and unbiased estimator DY of μ, $\mathbb{E}||DY - \mu||^2 \geq \mathbb{E}||\hat{\mu} - \mu||^2$.

Theorem 4.4. *We know that for any given linear functional ψ on M, there exists a unique vector $cv(\psi)$ in M, called the **coefficient vector of** ψ, such that $\psi(m) = \langle cv(\psi), m \rangle$ for all $m \in M$. Often the linear functional will be given initially in the form $\psi(m) = \langle x, m \rangle (m \in M)$ for some $x \in V$. Because $\langle x, m \rangle = \langle P_M x, m \rangle$ for all $m \in M$, we necessarily have $cv(\psi) = P_M x$ in this case. For ease of notation, it is convenient to define an inner product and norm for linear functionals on M as follows: $\langle \psi_1, \psi_2 \rangle = \langle cv(\psi_1), cv(\psi_2) \rangle, ||\psi|| = ||cv(\psi)||$.*

Definition 4.1. The **Gauss-Markov estimator (GME)**, $\hat{\psi}(Y)$, of a linear functional $\psi(\mu)$ of μ is defined by

$$\hat{\psi} = \hat{\psi}(Y) = \psi(P_M Y) = \langle cv(\psi), P_M Y \rangle = \langle cv(\psi), Y \rangle$$

Notice that for $x \in V$ the GME of the linear functional $\mu \to \langle x, \mu \rangle$ is

$$\langle P_M x, Y \rangle = \langle P_M x, P_M Y \rangle = \langle x, P_M Y \rangle$$

One must project either Y, or x, or both onto M before taking the inner product. In particular, when $x \in M, \langle x, Y \rangle$ is the GME of its expected value $\langle x, \mu \rangle$; this observation can frequently be used to obtain GMEs more or less at sight. To put it another way, if for a given linear functional ψ on M we can (aided by statistical intuition) guess at an $x \in M$ such that $\langle \mathbb{E}_\mu x, Y \rangle = \psi(\mu)$ for all $\mu \in M$, then $\hat{\psi}(Y) = \langle x, Y \rangle$.

Definition 4.2. Affine estimators

$T(Y)$ is called an affine estimator if $T(Y) = DY + d, D : V \to V, d \in V$ and D is a linear operator.

Definition 4.3. Risk function

The risk function of $T(Y)$ is defined as $R(Y,\mu) = \mathbb{E}||T(Y)-\mu||^2, \mu \in L = Im(X)$. This is also called the mean square error.

Proposition 4.2. Define $\mathscr{O} := \{T : \sup_{\mu \in L} R(T,\mu) < \infty\}$ where T is an affine estimator.

Then for any $T \in \mathscr{O}, R(T,\mu) \geq R(\hat{\mu},\mu), \mu \in L$.

Proof

Let $T(Y) = DY + d$, then $R(T,\mu) = \mathbb{E}||DY + d - \mu||^2 = \mathbb{E}||DY - D\mu + d + D\mu - \mu||^2$

$= \mathbb{E}||DY - D\mu||^2 + \underbrace{2\mathbb{E}\langle DY - D\mu, d + D\mu - \mu\rangle}_{0} + ||d + D\mu - \mu||^2.$

Therefore, $R(T,\mu) = \mathbb{E}||DY - D\mu||^2 + ||d + D\mu - \mu||^2$.

$\sup_{\mu \in L} R(T,\mu) < \infty \implies \sup_{\mu \in L}||d + D\mu - \mu||^2 < \infty \implies D\mu = \mu$. This is true since otherwise, there exists $\mu \in L, D\mu \neq \mu$, and $||d + t(D\mu - \mu)||^2$ which is a quadratic function for $t \in \mathbb{R}$ and it's not bounded in $\mathbb{R}$.

Hence $R(T,\mu) = \mathbb{E}||DY - \mu||^2 + ||d||^2 \geq \mathbb{E}||\hat{\mu} - \mu||^2 + ||d||^2$ where DY is an unbiased estimator.

So $R(T,\mu) \geq R(\hat{\mu},\mu), \mu \in L$. Moreover, $R(T,\mu) = R(\hat{\mu},\mu) \implies d = 0 \implies T(Y) = \hat{\mu}$.

□

Unbiased estimation of σ^2.

One candidate: RSS (Residue Sum of Squares) = $||Y - X\hat{\beta}||^2 = \sum_{i=1}^{\dim(V)}(Y_i - (X\hat{\beta})_i)^2$ where our model is $Y_j = (X\beta)_j + \xi_j$.

Proposition 4.3. Let $\tilde{\sigma^2} = \frac{||Y - X\hat{\beta}||^2}{\dim(V) - \dim(L)}$, then $\tilde{\sigma^2}$ is an unbiased estimator of σ^2.

Proof

$\mathbb{E}||Y - X\hat{\beta}||^2 = \mathbb{E}||Y - \hat{\mu}||^2 = \mathbb{E}||Y - P_L Y||^2 = \mathbb{E}||P_{L^\perp} Y||^2 = \mathbb{E}||P_{L^\perp}(\mu + \xi)||^2 = \mathbb{E}||P_{L^\perp}\xi||^2$ since $\mu \in L \implies P_{L^\perp}\mu = 0$.

So $\mathbb{E}||Y - X\hat{\beta}||^2 = tr(\Sigma_{P_{L^\perp}\xi}) = tr(P_{L^\perp}\Sigma_\xi P_{L^\perp}) = tr(P_{L^\perp}\sigma^2 I P_{L^\perp}) = \sigma^2 tr(P_{L^\perp}) = \sigma^2 \dim(P_{L^\perp}) = \sigma^2(\dim(V) - \dim(L))$.

□

Linear regression model with **normal noise**.

$\mathbf{Y} = \mathbf{X}\beta^* + \xi$

Basic assumption:

$Y \in V, \beta \in W$, inner product space.

$X : W \to V$, a linear operator. $Y \in V, \xi \in V$

$\sigma \sim N(0, \sigma^2 I)$.

$\mu = X\beta \in L = Im(X) \subset V$.

$Y \sim N(X\beta, \sigma^2 I_V), \mu \in L$.

Our **goal** is to estimate $\mu, \sigma^2(\beta, \sigma^2)$ based on Y.

One method is to try maximum likelihood estimators of μ, σ^2. We don't usually estimate β since it's not unique, hence we can not identify it.

$L(\mu, \sigma^2, y) = P_{\mu,\sigma^2}(y) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\exp(-\frac{1}{2\sigma^2}||y-\mu||^2), \mu \in L, \sigma^2 > 0$ and $\log L(\mu, \sigma^2, y) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}||y-\mu||^2$.

Then MLE is the defined as the $\operatorname{argmax}_{\mu\in L, \sigma^2>0} L(\mu, \sigma^2, y)$.

Proposition 4.4. The MLE for the linear model is $\hat{\mu} = P_L Y$ and $\hat{\sigma^2} = \frac{||Y - X\hat{\beta}||^2}{\dim(V)}$.

Proof

First find μ, then find σ^2.

(i) minimize $||y-\mu||^2$ with respect to $\mu \in L$, so $\hat{\mu} = P_L y$, the same as least square methods.

(ii) minimize $\frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}||y-\mu||^2$ with respect to $\sigma^2 > 0$.

$\frac{\partial}{\partial\sigma^2}\frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}||y-\mu||^2 = \frac{n}{2}\frac{1}{\sigma^2} - \frac{1}{2\sigma^2}||y-\mu||^2\frac{1}{\sigma^4} = 0 \implies \hat{\sigma^2} = \frac{||Y-\hat{\mu}||^2}{n} = \frac{||Y-X\hat{\beta}||^2}{\dim(V)}$ which is **biased**. $\square$

Proposition 4.5. Distribution of estimators

1. $\hat{\mu} \sim N(\mu, \sigma^2 P_L)$ normal distributions in V. Or $\hat{\mu} \sim N(\mu, \sigma^2 I_L)$ in space L.
2. $||\hat{\mu} - \mu||^2 \sim \sigma^2\chi^2_{\dim(L)}$.
3. $\hat{\mu}$ and $\hat{\sigma^2}$ are independent random variables.
4. $\hat{\sigma^2} \sim \frac{\sigma^2}{\dim(V)}\chi^2_{\dim(V)-\dim(L)}$.

Proof

Note that $Y \sim N(\mu, \sigma^2 I)$, $\hat{\mu} = P_L Y, \hat{\sigma^2} = \frac{||Y - X\hat{\beta}||^2}{\dim(V)}$. It's enough to prove that $\hat{\mu} = P_L Y$ and $Y - \hat{\mu} = Y - P_L Y = P_{L^\perp} Y$ are independent.

It's enough to check $P_L Y$ and $P_{L^\perp} Y$ are uncorrelated since they are normal.

$\Sigma_{P_L Y, P_{L^\perp} Y} = P_L \Sigma_Y P_{L^\perp} = \sigma^2 P_L P_{L^\perp} = 0$.

$\hat{\sigma^2} = \frac{||Y-\hat{\mu}||^2}{\dim V} = \frac{||Y - P_L Y||^2}{\dim V} = \frac{||P_{L^\perp} Y||^2}{\dim V} = \frac{||P_{L^\perp}(\mu+\xi)||^2}{\dim V} = \frac{||P_{L^\perp}\xi||^2}{\dim V} \sim \frac{\sigma^2}{\dim(V)} \chi^2_{\dim(V)-\dim(L)}$ since $\mu \in L$.

$\square$

Minimaxity of least square estimators: For a model $Y = X\beta + \xi, \xi \sim N(0, \sigma^2 I_V)$ in space V, $\beta \in W, X : W \to V \implies Y \sim N(\mu, \sigma^2 I_V), \mu = X\beta, \mu \in L = Im(X) \subset V$.

Then we will have the least square estimator $\hat{\beta} = \text{argmin}_{\beta \in W} ||Y - X\beta||^2$, then $\hat{\mu} = X\hat{\beta}$. And the risk is defined as $R(\mu, \hat{\mu}) = \mathbb{E}_\mu ||\hat{\mu} - \mu||^2 = \sigma^2 \dim(L)$ for any $\mu \in L$. Assume for now that σ^2 is known to us.

Definition 4.4. An estimator $T(X)$ is a minimax estimator for θ if $\sup_{\theta \in \Theta} R(\theta, T) = \inf_{\tilde{T}} \sup_{\theta \in \Theta} R(\theta, \tilde{T})$.

Reduction from minimax estimator to Bayes estimator: Suppose $X \sim P_\theta, \theta \in \Theta \subset V$ (inner product space with finite dimension). We will look at some prior distribution Π such that $\Pi(d\theta) = \pi(\theta) d\theta$ and $\pi(\theta)$ is called the prior density.

Definition 4.5. Bayes risk

For any estimator $T(X)$ of θ, define $R(\theta, T(X)) = \mathbb{E}_\theta ||T(X) - \theta||^2$ and the Bayes risk with respect to the prior Π as $R_\Pi(T) = \int_\Theta R(\theta, T) \Pi(d\theta) = \int_\Theta R(\theta, T) \pi(\theta) d\theta$.

Definition 4.6. The estimator $T_\Pi(X)$ is Bayes with respect to the prior Π if for any estimators $T(X)$, we have $R_\Pi(T) \geq R_\Pi(T_\Pi)$.

Proposition 4.6. Suppose there exists an estimator $T(X)$ and a sequence of prior Π_k distributions such that $R_{\Pi_k}(T_{\Pi_k}) \to \sup_{\theta \in \Theta} R(\theta, T)$ as $k \to \infty$ where $T_\Pi = \text{argmin}_T R_\Pi(T)$ is the Bayes estimator , then T is minimax.

Proof

For any estimator $\tilde{T}$, we have $\sup_{\theta \in \Theta}(R, \tilde{T}) \geq R_{\Pi_k}(\tilde{T}) \geq R_{\Pi_k}(T_{\Pi_k}) \to \sup_{\theta \in \Theta} R(\theta, T) \implies \sup_{\theta \in \Theta}(R, \tilde{T}) \geq$

$\sup_{\theta\in\Theta} R(\theta,T)$ since T_{Π_k} is Bayes for Π_k. Hence T is minimax. □

Definition 4.7. We have our prior $\Pi(d\theta) = \pi(\theta)d\theta$. Given θ, $X \sim P_\theta(dx) = P_\theta(x)dx$ where $P_\theta(x)$ is the density of X given θ. The posterior density is defined as $P(\theta|x) = \frac{P_\theta(x)\pi(\theta)}{\int_\Theta P_\theta(x)\pi(\theta)d\theta}$.

Proposition 4.7. If we define $T_\Pi(x) = \int \theta P(\theta|x)d\theta$ which is the posterior mean. Then $T_\Pi(x)$ is a Bayes estimator with respect to our prior Π.

Proof

Let $\tilde{\theta}$ be a random variable in Θ and $\tilde{\theta} \sim \Pi$. Given $\tilde{\theta} = \theta$, then $X \sim P(\cdot|\theta)$, and $(\tilde{\theta}, X)$ is a random couple in the space $\Theta \times S$ where space S is where X takes it values.

Note that $T_\Pi(x) = \int \theta P(\theta|x)d\theta = \mathbb{E}[\tilde{\theta}|x]$ where $\tilde{\theta}|x$ is the conditional density of $\tilde{\theta}$ given $X = x$. And $T_\Pi(X) = \mathbb{E}[\tilde{\theta}|x]$.

Plus, $R_\Pi(T) = \int_\Theta \mathbb{E}_\theta ||T(X)-\theta||^2\pi(\theta)d\theta = \int_\Theta \mathbb{E}(||T(X)-\tilde{\theta}||^2|\tilde{\theta}=\theta)\pi(\theta)d\theta = \mathbb{E}\mathbb{E}(||T(X)-\tilde{\theta}||^2|\tilde{\theta}) = \mathbb{E}||T(X)-\tilde{\theta}||^2$.

We have $R_\pi(T) = \mathbb{E}||T(X)-\tilde{\theta}||^2 = \mathbb{E}||T(X)-T_\Pi(X)+T_\Pi(X)-\tilde{\theta}||^2 = \mathbb{E}||T(X)-T_\Pi(X)||^2 + \mathbb{E}||T_\Pi(X)-\tilde{\theta}||^2 + 2\mathbb{E}\langle T(X)-T_\Pi(X), T_\Pi(X)-\tilde{\theta}\rangle$.

Now, $\mathbb{E}\langle T(X)-T_\Pi(X), T_\Pi(X)-\tilde{\theta}\rangle = \mathbb{E}_X\mathbb{E}(\langle T(X)-T_\Pi(X), T_\Pi(X)-\tilde{\theta}\rangle|X=x) = \mathbb{E}_X(\langle T(x)-T_\Pi(x), T_\Pi(x)-\mathbb{E}\tilde{\theta}|X=x\rangle)$ and $T_\Pi(x) - \mathbb{E}(\tilde{\theta}|X=x) = T_\Pi(x)-T_\Pi(x) = 0 \implies \mathbb{E}\langle T(X)-T_\Pi(X), T_\Pi(X)-\tilde{\theta}\rangle = 0 \implies R_\Pi(T) = \mathbb{E}||T(X)-T_\Pi(X)||^2 + \mathbb{E}||T_\Pi(X)-\tilde{\theta}||^2 \implies R_\Pi(T) \geq R_\Pi(T_\Pi)$.

□

Theorem 4.5. *Suppose $Y \sim N(\mu, \sigma^2 I_V), \mu \in L \subset V$ and $\hat{\mu} = P_L Y$. Then for any estimators $T(Y)$ of μ, we have $\sup_{\mu\in L}\mathbb{E}_\mu||T(Y)-\mu||^2 \geq \sup_{\mu\in L}\mathbb{E}_\mu||\hat{\mu}-\mu||^2$. In other words, $\sup_{\mu\in L} R(\mu,\hat{\mu}) = \inf_T \sup_{\mu\in L} R(\mu, T)$, or $\hat{\mu}$ is a minimax estimator of μ. And we assume $\dim(V) = n, \dim(L) = d$.*

Proof

Assume the prior distribution Π is $\mu \sim N(\theta, \tau^2 I_L), \theta \in L$ and $\tau^2 > 0$.

Note that the prior density $\pi(\mu) = \frac{1}{(\sqrt{2\pi}\tau)^d} e^{-\frac{|\mu-\theta||^2}{2\tau^2}}$ and the density of Y given μ is $P(y|\mu) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{||y-\mu||^2}{2\sigma^2}}$.

By the Bayes formula, $P(\mu|y)$ is proportional to $P(y|\mu)\pi(\mu) = Ce^{-\frac{|\mu-\theta||^2}{2\tau^2} - \frac{||y-\mu||^2}{2\sigma^2}}$ where C is a constant of μ which doesn't matter. Then, our guess is that $P(\mu|y) \sim N(a, b^2 I_L)$ which will give us $e^{-\frac{-||\mu-a||^2}{2b^2}}$.

We need $\frac{|\mu-\theta||^2}{\tau^2} + \frac{||y-\mu||^2}{\sigma^2} = \frac{||\mu-a||^2}{b^2}$ up to a constant does not depend on μ.

Or $(\frac{1}{\tau^2}+\frac{1}{\sigma^2})||\mu||^2 = \frac{1}{b}||\mu||^2$ and $\frac{1}{\tau^2}\langle\mu,\theta\rangle + \frac{1}{\sigma^2}\langle\mu,y\rangle = \frac{1}{b^2}\langle\mu,a\rangle$ for any $\mu\in L$. Note that $\mu\in\theta$, the only term that could be outside L is y so we should replace y to $P_L y$, or $\frac{1}{\tau^2}\langle\mu,\theta\rangle + \frac{1}{\sigma^2}\langle\mu,P_L y\rangle = \frac{1}{b^2}\langle\mu,a\rangle$ for any $\mu\in L$.

Therefore, $\langle\mu,\frac{1}{\tau^2}\theta+\frac{1}{\sigma^2}P_L y\rangle = \langle\mu,\frac{1}{b^2}a\rangle$ for any $\mu\in L \implies \frac{1}{\tau^2}\theta+\frac{1}{\sigma^2}P_L y = \frac{1}{b^2}a$.

So $\frac{1}{b^2}=\frac{1}{\tau^2}+\frac{1}{\sigma^2} \implies b^2=\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}$ and $\frac{1}{\tau^2}\theta+\frac{1}{\sigma^2}P_L y=\frac{1}{b^2}a \implies a=\frac{\sigma^2}{\sigma^2+\tau^2}\theta+\frac{\tau^2}{\sigma^2+\tau^2}P_L y$ and we can conclude that $\mu|Y=y\sim N(\frac{\sigma^2}{\sigma^2+\tau^2}\theta+\frac{\tau^2}{\sigma^2+\tau^2}P_L y, \frac{\tau^2\sigma^2}{\tau^2+\sigma^2}I_L)$.

The Bayes is given by posterior mean: $T_\Pi(Y)=\mathbb{E}[\tilde{\mu}|Y]=\frac{\sigma^2}{\sigma^2+\tau^2}\theta+\frac{\tau^2}{\sigma^2+\tau^2}\underbrace{P_L y}_{\hat{\mu}}$.

To prove minimaxity of $\hat{\mu}$, we can choose $\Pi_k\sim N(0,\tau_k^2 I_L)$ where $\tau_k^2\to\infty$, then $T_{\Pi_k}=\frac{\tau_k^2}{\tau_k^2+\sigma^2}\hat{\mu}$.

First, let's consider the risk $R(\mu,T_{\Pi_k})=\mathbb{E}_\mu||T_{\Pi_k}(y)-\mu||^2=\mathbb{E}_\mu||\frac{\tau_k^2}{\tau_k^2+\sigma^2}\hat{\mu}-\mu||^2=\mathbb{E}_\mu||\frac{\tau_k^2}{\tau_k^2+\sigma^2}(\hat{\mu}-\mu)-\frac{\sigma^2}{\tau_k^2+\sigma^2}\mu||^2$.

Which is equal to $\mathbb{E}_\mu(\frac{\tau_k^2}{\tau_k^2+\sigma^2})^2(\hat{\mu}-\mu)^2+(\frac{\sigma^2}{\tau_k^2+\sigma^2})^2||\mu||^2-\underbrace{2\mathbb{E}_\mu\langle\frac{\tau_k^2}{\tau_k^2+\sigma^2}(\hat{\mu}-\mu),\frac{\sigma^2}{\tau_k^2+\sigma^2}\mu\rangle}_{=2\langle\mathbb{E}_\mu\frac{\tau_k^2}{\tau_k^2+\sigma^2}(\hat{\mu}-\mu),\frac{\sigma^2}{\tau_k^2+\sigma^2}\mu\rangle=0}$.

So $R(\mu,T_{\tau k})=\mathbb{E}_\mu(\frac{\tau_k^2}{\tau_k^2+\sigma^2})^2(\hat{\mu}-\mu)^2+(\frac{\sigma^2}{\tau_k^2+\sigma^2})^2||\mu||^2=(\frac{\tau_k^2}{\tau_k^2+\sigma^2})^2\sigma^2\dim(L)+(\frac{\sigma^2}{\tau_k^2+\sigma^2})^2||\mu||^2$ since $P_L(y-\mu)=P_L(\mu+\xi)-\mu=P_L\xi$.

So $R_{\Pi_k}(T_k)=(\frac{\tau_k^2}{\tau_k^2+\sigma^2})^2\sigma^2\dim(L)+(\frac{\sigma^2}{\tau_k^2+\sigma^2})^2\underbrace{\int_L||\mu||^2\Pi_k(d\mu)}_{\mathbb{E}||\mu||^2=\tau_k^2 d}=(\frac{\tau_k^2}{\tau_k^2+\sigma^2})^2\sigma^2 d+(\frac{\sigma^2}{\tau_k^2+\sigma^2})^2\tau_k^2 d$

And $\lim_{\tau_k^2\to\infty}R_{\Pi_k}(T_k)=\sigma^2 d=\sup_{\mu\in L}\mathbb{E}_\mu||\hat{\mu}-\mu||^2$.

It follows that $\hat{\mu}$ is minimax by proposition 4.6.

□

Remark: It's not hard to shown that for a proper prior distribution, the Bayes estimator will be biased. Since $\hat{\mu}$ is unbiased, we don't expect it to be a Bayes estimator.

Proposition 4.8. Let's go back to our model, $Y=X\beta+\xi$ where $\mathbb{E}\xi=0$ and $\Sigma_\xi=\sigma^2 I_V$ and $\mu=X\beta=\mathbb{E}Y\in Im(X)=V\subset V$, $d=\dim(L)$.

Let P_{L,σ_0^2} be the family of distributions P satisfying the model $\mu=\mu(P)$ and $\sigma^2=\sigma^2(P)\le\sigma_0^2$. Or we are bounding the variance.

Let $T(Y)$ be an estimator of $\mu=\mu(P)$, and the risk $R(P,T)=\mathbb{E}_P||T(Y)-\mu(P)||^2$. And $R(P,\hat{\mu})=\mathbb{E}_P||\hat{\mu}-\mu(P)||^2=\sigma(P)^2 d$. Also, $\sup_{P\in P_{L,\sigma_0^2}}R(P,\hat{\mu})=\sigma_0^2 d$ since we are taking sup on both sides.

For any estimator $T(Y)$, $\sup_{P\in P_{L,\sigma_0^2}}\mathbb{E}_P||T(Y)-P||^2\ge\sigma_0^2 d=\sup_{P\in P_{L,\sigma_0^2}}R(P,\hat{\mu})$. It follows that $\hat{\mu}$ is minimax.

Proof

Consider $N_{L,\sigma_0^2} = \{N(\mu, \sigma_0^2 I_L), \mu \in L\}$, then clearly $N_{L,\sigma_0^2} \subset P_{L,\sigma_0^2}$.

We can write down the following

$$\sup_{P \in P_{L,\sigma_0^2}} \mathbb{E}_P ||\hat{\mu} - \mu(P)||^2 = \sigma_0^2 d = \sup_{P \in N_{L,\sigma_0^2}} \mathbb{E}_P ||\hat{\mu} - \mu(P)||^2 \leq \sup_{P \in N_{L,\sigma_0^2}} \mathbb{E}_P ||T(Y) - \mu(P)||^2 \leq \sup_{P \in P_{L,\sigma_0^2}} \mathbb{E}_P ||T(Y) - \mu(P)||^2$$

which is true for any estimator $T(Y)$.

Therefore, for any estimator $T(Y)$, $\sup_{P \in P_{L,\sigma_0^2}} R(P, L) \geq \sup_{P \in P_{L,\sigma_0^2}} R(P, \hat{\mu})$. It follows that $\hat{\mu}$ is minimax. $\square$

Definition 4.8. An estimator $T(\tilde{Y})$ is admissible if there is no other $T(Y)$ such that T improves $\tilde{T}$, or no T such that $\mathbb{E}_\mu ||T(Y) - \mu||^2 \leq \mathbb{E}_\mu ||T(\tilde{Y}) - \mu||^2, \mu \in L$ with strict inequality for some μ.

Theorem 4.6. *If T_Π is a* ***unique*** *Bayes estimator for some prior Π, then T_Π is admissible.*

Proof

If there exists $T(Y)$ such that $\mathbb{E}_\mu ||T(Y) - \mu||^2 \leq \mathbb{E}_\mu ||T_\Pi(Y) - \mu||^2, \mu \in L$, this imply that $R_\Pi(T) \leq R_\Pi(T_\Pi)$, this means $R_\Pi(T) = R_\Pi(T_\Pi)$ since T_Π is Bayes, so this imply that $T = T_\Pi$. So there's no better estimator than T_Π. $\square$

Proposition 4.9. Stein's Identity

Assume $X \sim N(\theta, \sigma^2 I_d), \theta \in \mathbb{R}^d$, let g be a smooth function $\mathbb{R}^d \to \mathbb{R}^d$, then $\mathbb{E}_\theta \langle X - \theta, g(X) \rangle = \sigma^2 \mathbb{E}_\theta \text{div}(g(X))$ where $\text{div}(g(X)) = \frac{\partial g_1(X)}{\partial X_1} + \cdots + \frac{\partial g_d(X)}{\partial X_d}$

Proof

For $d = 1$, we need to prove $\mathbb{E}_\theta (X - \theta) g(X) = \sigma^2 \mathbb{E}_\theta g'(X)$ which can be verified by integral by parts.

The left-hand side is

$$\mathbb{E}_\theta [g(X)(X - \theta)] = \frac{1}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}} g(x)(x - \theta) e^{-\frac{(x-\theta)^2}{2\sigma^2}} dx$$

Use integration by parts with $u = g(x)$ and $dv = (x - \theta) e^{-\frac{(x-\theta)^2}{2\sigma^2}} dx$ to get

$$\mathbb{E}_\theta[g(X)(X-\theta)] = \frac{1}{\sqrt{2\pi}\sigma}[-\sigma^2 g(x)e^{-\frac{(x-\theta)^2}{2\sigma^2}}]|_{-\infty}^{\infty} + \sigma^2 \int_{\mathbb{R}} g'(x)e^{-\frac{(x-\theta)^2}{2\sigma^2}}\, dx$$

The condition on g' is enough to ensure that the first term is 0 and what remains on the right-hand side is $\sigma^2\mathbb{E}_\theta[g'(X)]$.

For $d > 1$, we need to prove $\sum_{i=1}^d \mathbb{E}_\theta(X_i-\theta_i)g_i(X) = \sigma^2 \sum_{i=1}^d \mathbb{E}_\theta \frac{\partial g_i(X)}{\partial X_i}$, which can be proved by using previous results and condition on coordinates. □

Definition 4.9. If $u:\mathbb{R}^n \to \mathbb{R}$, then the gradient is defined as $\nabla u = (\frac{\partial u}{\partial x_1}, \cdots, \frac{\partial u}{\partial x_n})$. Note that $\mathrm{div}(\nabla u) = \Delta u = \sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2}$ is the Laplacian operator.

Theorem 4.7. *Stein's theorem*

For the model $Y = X\beta + \xi$ where $\mathbb{E}\xi = 0$ and $\Sigma_\xi = \sigma^2 I_V$ and $\mu = X\beta = \mathbb{E}Y \in Im(X) = L \subset V$, $d = \dim(L)$, and $\hat{\mu} = P_L Y$.

If $\dim(L) \geq 3$, then there exists a estimator $T(Y)$ of μ such that for any $\mu \in L$, $\mathbb{E}_\mu||T(Y)-\mu||^2 < \mathbb{E}_\mu||\hat{\mu}-\mu||^2 = \sigma^2\dim(L)$. Or $\hat{\mu}$ is an inadmissible estimator.

Proof

We can construct $T(Y) = \hat{\mu} + \sigma^2 g(\hat{\mu})$ where $g: L \to L$ is a smooth function, we can always identify L with $\mathbb{R}^d$ by choosing coordinates.

Now, $\mathbb{E}_\mu||T(Y)-\mu||^2 = \mathbb{E}_\mu||\hat{\mu}-\mu+\sigma^2 g(\hat{\mu})||^2 = \mathbb{E}_\mu||\hat{\mu}-\mu||^2 + 2\sigma^2\mathbb{E}_\mu\langle\hat{\mu}-\mu, g(\hat{\mu})\rangle + \sigma^4\mathbb{E}_\mu||g(\hat{\mu})||^2$ so we need $\mathbb{E}_\mu\langle\hat{\mu}-\mu, g(\hat{\mu})\rangle$ to be negative to reduce the loss.

We have $\hat{\mu} = P_L Y \sim N(\mu, \sigma^2 I_L)$, by Stein's identity, we have $\mathbb{E}_\mu\langle\hat{\mu}-\mu, g(\hat{\mu})\rangle = \sigma^2\mathbb{E}_\mu \mathrm{div}(g(\hat{\mu}))$, so $\mathbb{E}_\mu||T(Y)-\mu||^2 = \mathbb{E}_\mu||\hat{\mu}-\mu||^2 + 2\sigma^4\mathbb{E}_\mu \mathrm{div}(g(\hat{\mu})) + \sigma^4\mathbb{E}_\mu||g(\hat{\mu})||^2$.

Let's assume $L = \mathbb{R}^d$ by coordinates for simplicity, and we will choose $g(x) = \nabla \log\psi(x), x \in \mathbb{R}^d$, where $\psi: \mathbb{R}^d \to \mathbb{R}$ $\psi(x) > 0$ and $\psi(x)$ is smooth and ψ is not a constant. Note that $g(x) = \nabla\log\psi(x) = \frac{\nabla\psi(x)}{\psi(x)}$, so $\mathrm{div}(g(x)) = \frac{\Delta\psi(x)\cdot\psi(x) - ||\nabla\psi||^2}{\psi^2(x)}$. As a result, $\mathrm{div}(g(x)) = \frac{\Delta\psi(x)}{\psi(x)} - \underbrace{\frac{||\nabla\psi(x)||^2}{\psi^2(x)}}_{||g(x)||^2} = \frac{\Delta\psi(x)}{\psi(x)} - ||g(x)||^2$.

Then $\mathbb{E}_\mu||T(Y)-\mu||^2 = \mathbb{E}_\mu||\hat{\mu}-\mu||^2 + 2\sigma^4\mathbb{E}_\mu\frac{\Delta\psi(\hat{\mu})}{\psi(\hat{\mu})} - 2\sigma^4\mathbb{E}_\mu||g(\hat{\mu})||^2 + \sigma^4\mathbb{E}_\mu||g(\hat{\mu})||^2$.

So $\mathbb{E}_\mu||T(Y)-\mu||^2 = \mathbb{E}_\mu||\hat{\mu}-\mu||^2 + 2\sigma^4\mathbb{E}_\mu\frac{\Delta\psi(\hat{\mu})}{\psi(\hat{\mu})} - \sigma^4\mathbb{E}_\mu||g(\hat{\mu})||^2$. Next we should choose a harmonic function ψ to make $\Delta\psi(\hat{\mu}) = 0$ to improve the risk.

We need to have $\psi > 0$, ψ is not constant, ψ is smooth and ψ is harmonic, such functions only exist for $d \geq 3$, this is a **fact**.

For such choice of ψ and $g = \nabla\log\psi$, we have $\mathbb{E}_\mu||T(Y)-\mu||^2 = \mathbb{E}_\mu||\hat{\mu}-\mu||^2 - \sigma^4\mathbb{E}_\mu||g(\hat{\mu})||^2$.

To prove that $\mathbb{E}_\mu||T(Y)-\mu||^2 < \mathbb{E}_\mu||\hat{\mu}-\mu||^2, \mu \in L = \mathbb{R}^d$, it remains to show $\mathbb{E}_\mu||g(\hat{\mu})||^2 > 0$.

Since $g \neq 0$, and g is continuous, there exists $x_0 \in \mathbb{R}^d$ and a small $\delta > 0$ such that $||g(x)||^2 \geq c > 0$ for all $x \in B(x_0,\delta)$, the ball centered at x_0 with radius δ. It follows that $\mathbb{E}_\mu||g(\hat{\mu})||^2 \geq c\mathbb{P}_\mu\{\hat{\mu} \in B(x_0,\delta)\} > 0$ and $\mathbb{P}_\mu\{\hat{\mu} \in B(x_0,\delta)\} > 0$ since $\hat{\mu}$ follows a nonsingular normal distribution on $L = \mathbb{R}^d$.

A choice of ψ can be $\psi(x) = ||x||^{-(d-2)}$ for $d \geq 3$ which is a potential field function and ψ is a harmonic function. Note ψ is not defined at 0, which is not a big trouble. Note that $\psi > 0, \psi$ is not a constant. When $d < 3$, there will be change of signs and therefore ψ doesn't exist, a formal proof can be seen in some mathematical physics textbooks.

Now, $g(x) = \nabla \log \psi(x) = \frac{\nabla\psi(x)}{\psi(x)}$. Note that $\nabla\psi(x) = \nabla(||x||^2)^{1-\frac{d}{2}} = (1-\frac{d}{2})(||x||^2)^{-\frac{d}{2}}\underbrace{\nabla||x||^2}_{2x} = (2-d)\frac{x}{||x||^d}$

and $g(x) = \frac{\nabla\psi(x)}{\psi(x)} = \frac{(2-d)\frac{x}{||x||^d}}{\frac{1}{||x||^{d-2}}} = (2-d)\frac{x}{||x||^2}$.

And $T(Y) = \hat{\mu} + \sigma^2 g(\hat{\mu}) = \hat{\mu} - \sigma^2(d-2)\frac{\hat{\mu}}{||\hat{\mu}||^2} = \hat{\mu}(1-\frac{\sigma^2(d-2)}{||\hat{\mu}||^2})$ which is called **James-Stein estimator**. It can also be constructed based on the Bayesian approach. Note that $\mathbb{E}_\mu||T(Y)-\mu||^2 = \mathbb{E}_\mu||\hat{\mu}-\mu||^2 - \sigma^4\mathbb{E}_\mu||g(\hat{\mu})||^2 = \mathbb{E}_\mu||T(Y)-\mu||^2 = \underbrace{\mathbb{E}_\mu||\hat{\mu}-\mu||^2}_{\sigma^2 d} - \sigma^4\mathbb{E}_\mu(2-d)^2\frac{||\hat{\mu}||^2}{||\hat{\mu}||^4} = \sigma^2 d - \sigma^4(d-2)^2\mathbb{E}_\mu\frac{1}{||\hat{\mu}||^2}$ since $g(x) = (2-d)\frac{x}{||x||^2}$.

Now $\hat{\mu} \sim N(\mu,\sigma^2) \implies ||\hat{\mu}||^2 \sim \sigma^2\chi^2_{d,\frac{||\mu||}{\sigma}}$, we have $\mathbb{E}_\mu||T(Y)-\mu||^2 = \sigma^2 d - \sigma^2(d-2)^2\mathbb{E}\frac{1}{\chi^2_{d,\frac{||\mu||}{\sigma}}}$.

From previous lecture notes, we have

$$\chi^2_{d,\frac{||\mu||}{\sigma}} = \sum_{k=0}^{\infty} e^{-\frac{||\mu||^2}{2\sigma^2}}\frac{(\frac{||\mu||^2}{2\sigma^2})^k}{k!}\chi^2_{d+2k} \implies \mathbb{E}\frac{1}{\chi^2_{d,\frac{||\mu||}{\sigma}}} = \sum_{k=0}^{\infty} e^{-\frac{||\mu||^2}{2\sigma^2}}\frac{(\frac{||\mu||^2}{2\sigma^2})^k}{k!}\underbrace{\mathbb{E}\frac{1}{\chi^2_{d+2k}}}_{\frac{1}{d-2+2k}}$$

where the last equation is true since chi-square distribution is a special case of Gamma distribution.

So we will end up with $\mathbb{E}\frac{1}{\chi^2_{d,\frac{||\mu||}{\sigma}}} = \sum_{k=0}^{\infty} e^{-\lambda}\frac{k}{k!}\frac{1}{d-2+2k} = \mathbb{E}_{\nu\sim\text{Poisson}(\lambda)}\frac{1}{d-2+2\nu}$ where $\lambda = \frac{||\mu||^2}{2\sigma^2}$.

So $\mathbb{E}_\mu||T(Y)-\mu||^2 = \sigma^2 d - \sigma^2(d-2)^2\mathbb{E}_{\nu\sim\text{Poisson}(\lambda)}\frac{1}{d-2+2\nu}$ where $\lambda = \frac{||\mu||^2}{2\sigma^2}$.

For $\mu = 0$, $\mathbb{E}_\mu||T(Y)-\mu||^2 = 2\sigma^2$, a great reduction for the variance from $d\sigma^2$ to $2\sigma^2$.

□

Orthogonal designs

Let $Y = X\beta + \xi, y \in \mathbb{R}^n, \xi \sim N(0,\sigma^2 I_n), \beta \in \mathbb{R}^p, X$ is a $n \times p$ matrix, also called the design matrix.

We can write as $x_1,\cdots,x_p \in \mathbb{R}^n$, and $Y = \beta_1 x_1 + \cdots + \beta_p x_p + \xi$. Then the least-square estimator is $\hat{\beta} = \text{argmin}_{\beta\in\mathbb{R}^p}||Y - X\beta||^2$ and $X\hat{\beta} = P_L Y, L = Im(X) \subset \mathbb{R}^n$.

When X^TX is not singular, there is a unique solution by the normal equation, and $\hat{\beta} = (X^TX)^{-1}X^TY$.

We have $(X^TX)_{ij} = \sum_{k=1}^n X^T_{ik}X_{kj} = \sum_{k=1}^n X_{ki}X_{kj} = \langle x_i, x_j\rangle$. Or $X^TX = (\langle x_i,x_j\rangle)_p$ where we call this matrix **Gram**

matrix.

Note that X^TX is positive semidefinite since the quadratic form is $\sum_{i,j}\langle x_i,x_j\rangle c_ic_j = \langle \sum_i c_ix_i, \sum_j c_jx_j\rangle = ||\sum_i c_ix_i||^2 \geq 0$. If we assume that $x_1,\cdots,x_p$ are independent (or $\sum_i c_ix_i = 0 \implies c_i = 0$), then $||\sum_i c_ix_i||^2 = 0 \implies c_i = 0$, which is equivalent to that the Gram matrix X^TX is positive definite.

Proposition 4.10. If we have a nonsingular $p\times p$ matrix A, or $\det(A)\neq 0$. Let A has entries a_{ij}, so we can take the i th row and j th column. Note that $\tilde{A}$ has entries $\tilde{a_{ij}} = (-1)^{i+j}\det(\tilde{A_{ij}})$ where $\tilde{A_{ij}}$ is the minor. Then $A^{-1} = \frac{\tilde{A}^T}{\det A}$.

Theorem 4.8. *Hotelling's Theorem*

Let $\hat\beta = (\hat\beta_1,\cdots,\hat\beta_p)$, suppose X^TX is nonsingular, then for any $j = 1,\cdots,p$, $\mathbb{V}\mathrm{ar}(\hat\beta_j) \geq \frac{\sigma^2}{||x_j||^2}$. Moreover, if $\mathbb{V}\mathrm{ar}(\hat\beta_j) = \frac{\sigma^2}{||x_j||^2} \implies x_j \perp x_i$ for $i\neq j$.

Proof

Consider the Covariance $\Sigma_{\hat\beta} = (X^TX)^{-1}X^T\Sigma_Y X(X^TX)^{-1} = (X^TX)^{-1}X^T\sigma^2 I_n X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$.

Then $\mathbb{V}\mathrm{ar}(\hat\beta_j) = \langle \Sigma_{\hat\beta}e_j, e_j\rangle = \sigma^2\langle (X^TX)^{-1}e_j, e_j\rangle = \sigma^2(X^TX)^{-1}_{jj}$ where e_j is the canonical basis of $\mathbb{R}^n$.

Without loss of generality, we can take $j = 1$, then

$$X^TX = \begin{bmatrix} \langle x_1,x_1\rangle & b^T \\ b & C \end{bmatrix}$$

where $b\in\mathbb{R}^{p-1}$ with entries $b_j = \langle x_1,x_j\rangle, 2\leq j\leq n$ and C is a $(p-1)\times(p-1)$ Gram matrix with entries $C_{ij} = \langle x_i,x_j\rangle, 2\leq i,j\leq p$.

Then $(X^TX)^{-1}_{11} = \frac{\det(C)}{\det(X^TX)}$.

And

$$\det(X^TX) = \det\left(\begin{bmatrix} \langle x_1,x_1\rangle & b^T \\ b & C \end{bmatrix}\right)\underbrace{\det\left(\begin{bmatrix} 1 & 0 \\ -C^{-1}b & I_{p-1} \end{bmatrix}\right)}_{1} = \det\left(\begin{bmatrix} \langle x_1,x_1\rangle & b^T \\ b & C \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -C^{-1}b & I_{p-1} \end{bmatrix}\right)$$

Note that

$$\begin{bmatrix} \langle x_1,x_1\rangle & b^T \\ b & C \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -C^{-1}b & I_{p-1} \end{bmatrix} = \begin{bmatrix} \langle x_1,x_1\rangle - b^TC^{-1}b & b^T \\ \underbrace{0}_{b-CC^{-1}b} & C \end{bmatrix}$$

Use minor decomposition,

$$\det(\begin{bmatrix} \langle x_1, x_1 \rangle - b^T C^{-1} b & b^T \\ 0 & C \end{bmatrix}) = (\langle x_1, x_1 \rangle - \langle C^{-1}b, b \rangle) \det(C)$$

So $\mathbb{V}\text{ar}(\hat{\beta}_1) = \sigma^2 (X^T X)^{-1} = \sigma^2 \frac{\det(C)}{\det(X^T X)} = \sigma^2 \frac{1}{\langle x_1, x_1 \rangle - \langle C^{-1}b, b \rangle}$.

It follows that $\mathbb{V}\text{ar}(\hat{\beta}_j) = \frac{\sigma^2}{||x_j||^2 - \langle C^{-1}b, b \rangle}$. Note that C and b depend on j, but for simplicity we don't use the subscripts for now. And C is a $(p-1) \times (p-1)$ a Gram matrix, positive semidefinite, and C is nonsingular.

If a matrix is positive definite, then by definition, its smaller part is also positive definite. So C^{-1} exists and is positive definite and $\langle C^{-1}b, b \rangle > 0$ for $b \neq 0$.

So $\mathbb{V}\text{ar}(\hat{\beta}_1) \geq \frac{\sigma^2}{||x_1||^2}$ and $\mathbb{V}\text{ar}(\hat{\beta}_1) = \frac{\sigma^2}{||x_1||^2} \implies \langle C^{-1}b, b \rangle = 0 \implies \langle C^{-1/2}b, C^{-1/2}b \rangle = 0 \implies C^{-1/2}b = 0 \implies b = 0 \implies \langle x_1, x_j \rangle = 0, \forall j$.

$\square$

Let $\mathscr{D}_{c_1, \cdots, c_p}$ be the set of $n \times p$ design matrix X such that $X^T X$ is nonsingular and $||x_j||^2 = c_j^2$, then the variance of least square estimator, $\hat{\beta}_j$ are minimized for the design X such that $x_i \perp x_j, i \neq j$. We call this orthogonal design. In this case, $X^T X$ becomes a diagonal matrix with entries $c_1^2 \cdots c_p^2$ and $\mathbb{V}\text{ar}(\hat{\beta}_j) = \frac{\sigma^2}{c_j^2}$.

Suppose $Y \sim N(\mu, \sigma^2 I_V)$ in V, $\mu \in L \subset V$, $\dim(V) = n, \dim(L) = d$, say $L = Im(X), X : W \to V$.

Let $L_0 \subset L$, we want to test the hypothesis H_0 against H_a where $H_0 : \mu \in L_0$ and $H_a : \mu \notin L_0$. And we will use the **likelihood ratio test**.

Definition 4.10. Likelihood Ratio Test

The likelihood ratio is $\Lambda = \frac{\sup_{\mu \in L, \sigma^2 > 0} L(\mu, \sigma^2, y)}{\sup_{\mu \in L_0, \sigma^2 > 0} L(\mu, \sigma^2, y)}$ where we don't care about σ^2. And we reject H_0 if $\Lambda \geq c$ and we don't reject H_0 if $\Lambda < c$. We need to choose c so that under H_0, or $\mu \in L_0$, the probability to reject H_0, or $\mathbb{P}_{\mu, \sigma^2}\{\Lambda \geq c\} = \alpha$ for any $\mu \in L_0, \sigma^2 > 0$. And α is also called the significance level of the test. Generally, it's not possible to satisfy this equation since μ is arbitrary.

Recall the likelihood function is $L(\mu, \sigma^2, y) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp(-\frac{||y-\mu||^2}{2\sigma^2})$. And the maximum likelihood estimator for the whole model is $(\hat{\mu}, \hat{\sigma^2}) = \text{argmax}_{\mu \in L, \sigma^2 > 0} L(\mu, \sigma^2, y) = (P_L y, \frac{||y - P_L y||^2}{n})$.

Similarly, we can write the maximum likelihood estimator for H_0, is $(\hat{\mu_0}, \hat{\sigma_0^2}) = \text{argmax}_{\mu \in L_0, \sigma^2 > 0} L(\mu, \sigma^2, y) = (P_{L_0} y, \frac{||y - P_{L_0} y||^2}{n})$.

Note that $L(\hat{\mu}, \hat{\sigma^2}, y) = \frac{1}{(2\pi)^{n/2}(\hat{\sigma^2})^{n/2}} \exp(-\frac{||Y - P_L y||^2}{2\hat{\sigma^2}}) = \frac{1}{(2\pi)^{n/2}(\hat{\sigma^2})^{n/2}} \exp(-\frac{n}{2})$ and $L(\hat{\mu_0}, \hat{\sigma_0^2}, y) = \frac{1}{(2\pi)^{n/2}(\hat{\sigma_0^2})^{n/2}} \exp(-\frac{n}{2})$

And $\Lambda = \frac{L(\hat{\mu}, \hat{\sigma^2}, y)}{L(\hat{\mu_0}, \hat{\sigma_0^2}, y)} = (\frac{\hat{\sigma_0^2}}{\hat{\sigma^2}})^{n/2}$.

The likelihood ratio test is given as $\Lambda \geq c \Leftrightarrow (\frac{\hat{\sigma_0^2}}{\hat{\sigma^2}})^{n/2} \geq c \Leftrightarrow \frac{\hat{\sigma_0^2}}{\hat{\sigma^2}} \geq c' \Leftrightarrow \frac{||y - P_{L_0} y||^2}{||y - P_L y||^2} = \frac{||y - P_L y||^2 + ||P_L y - P_{L_0} y||^2}{||y - P_L y||^2} = 1 + \frac{||P_L y - P_{L_0} y||^2}{||y - P_L y||^2} \geq c' \Leftrightarrow \frac{||P_L y - P_{L_0} y||^2}{||y - P_L y||^2} \geq c''$.

Now we can consider the statistic $T = \frac{||P_L Y - P_{L_0} Y||^2}{||Y - P_L Y||^2}$ and we reject H_0 if $T \geq c$. Note that $Y - P_L Y \perp P_L Y - P_{L_0} Y$ so they are uncorrelated and independent since they are normal. $||Y - P_L Y||^2 = ||P_{L^\perp} Y||^2 \sim \sigma^2 \chi^2_{n-d}$ since $L^\perp \mu = 0$ and $||\underbrace{P_L Y - P_{L_0} Y}_{P_{L-L_0} Y}||^2 \sim \sigma^2 \chi^2_{d-d_0, \frac{||(P_L - P_{L_0})\mu||}{\sigma}}$.

Since $T \sim \frac{\chi^2_{d-d_0, \frac{||(P_L - P_{L_0})\mu||}{\sigma}}}{\chi^2_{n-d}} \sim \mathscr{F}_{d-d_0, n-d, \frac{||(P_L - P_{L_0})\mu||}{\sigma}}$, we have the $\mathscr{F}$ test.

Under H_0, $\mu \in L_0, (P_L - P_{L_0})\mu = 0$. It follows that $T \sim \mathscr{F}_{d-d_0, n-d}$ and we have a parameter that's parameter-free with respect to μ and σ^2.

So $\mathbb{P}_{\mu, \sigma^2}\{T \geq c\} = \mathbb{P}\{\mathscr{F}_{d-d_0, n-d} \geq c\} = \alpha$. Then we reject H_0 if $T \geq c(d - d_0, n - d)$ and we don't reject otherwise. To compute the power of the test, we need the non-central parameter $\frac{||(P_L - P_{L_0})\mu||}{\sigma}$.

Theorem 4.9. *Gram Schmidt orthogonalization*

Let V be a vector space with an inner product. Suppose $x_1, x_2, \cdots, x_n$ is a basis for V, and

$v_1 = x_1$, *then normalize it*

$v_2 = x_2 - \frac{\langle x_2, v_1 \rangle}{\langle v_1, v_1 \rangle} v_1$, *then normalize it*

$v_3 = x_3 - \frac{\langle x_3, v_1 \rangle}{\langle v_1, v_1 \rangle} v_1 - \frac{\langle x_3, v_2 \rangle}{\langle v_2, v_2 \rangle} v_2$, *then normalize it*

$\ldots$

$v_n = x_n - \frac{\langle x_n, v_1 \rangle}{\langle v_1, v_1 \rangle} v_1 - \cdots - \frac{\langle x_n, v_{n-1} \rangle}{\langle v_{n-1}, v_{n-1} \rangle} v_{n-1}$, *then normalize it*

Then $v_1, v_2, \cdots, v_n$ is an orthonormal basis for V.

Example 4.1. Consider now the simple linear models, $Y_i = \beta_0 + \beta_1 X_i + \xi_i, i = 1, \cdots, n$ and ξ are iid $N(0, \sigma^2)$. Or we can write $Y = X\beta + \xi$. Where $Y \in \mathbb{R}^n, \beta \in \mathbb{R}^2, \xi \in \mathbb{R}^n \sim N(0, \sigma^2 I_n)$.

Let $\mathbf{1} \in \mathbb{R}^n, x \in \mathbb{R}^n$, we have $Y = \beta_0 \mathbf{1} + \beta_1 x + \xi$ and $L =$ linear span$(\mathbf{1}, x)$ and $L_0 =$ linear span$(\mathbf{1})$.

By **Gram Schmidt orthogonalization**: $e_1 = \frac{1}{\sqrt{n}} \mathbf{1}$ and $||e_1|| = 1$. Also $e_2 = \frac{x - \langle x, e_1 \rangle e_1}{||x - \langle x, e_1 \rangle e_1||}$ and $||e_2|| = 1, e_1 \perp e_2$.

Now, $\langle x, e_1 \rangle e_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i \frac{1}{\sqrt{n}} \mathbf{1} = \bar{x}\mathbf{1}$ and $||x - \langle x, e_1 \rangle e_1||^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = nS_x^2$ and $e_2 = \frac{x - \bar{x}\mathbf{1}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$.

Also $P_L Y = \underbrace{\langle Y, e_1 \rangle}_{c_1} e_1 + \underbrace{\langle Y, e_2 \rangle}_{c_2} e_2$.

Note that $\beta_0 \mathbf{1} + \beta_1 x = (\beta_0 + \beta_1 \bar{x})\mathbf{1} + \beta_1 (x - \bar{x}\mathbf{1}) = (\beta_0 + \beta_1 \bar{x})\sqrt{n} e_1 + \beta_1 \sqrt{n} S_x e_2$.

Therefore, $c_1 = (\beta_0 + \beta_1 \bar{x})\sqrt{n}$ and $c_2 = \beta_1 \sqrt{n} S_x$, so $\beta_1 = \frac{c_2}{\sqrt{n} S_x}$ and $\beta_0 = \frac{c_1}{\sqrt{n}} - \beta_1 \bar{x}$.

It follows that $\hat{c_1} = \langle \hat{Y}, e_1 \rangle = \sqrt{n}\bar{Y}$ and $\hat{c_2} = \langle \hat{Y}, e_2 \rangle = \langle Y - \bar{Y}\mathbf{1}, e_2 \rangle = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{n} S_x}$ since e_1 and e_2 are orthogonal.

Hence $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{n} S_x}$ and $\hat{\rho} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_x S_y}$ is called the sample correlation coefficient so $\hat{\beta}_1 = \frac{\rho S_x S_y}{S_x^2} = \hat{\rho} \frac{S_y}{S_x}$.

And $\beta_1 = \hat{\rho}\frac{S_y}{S_x}$, $\beta_0 = \bar{Y} - \hat{\beta}_1\bar{x}$.

We can write $P_L Y = \hat{\beta}_0\mathbf{1} + \hat{\beta}_1 x$ and $P_{L_0}y = \bar{Y}$ and we can then use the F-test.

Example 4.2. The least-square estimator for $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ is $\hat{\beta}_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$ and $\hat{\beta}_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$ and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}_1 - \hat{\beta}_2\bar{X}_2$.

Two-way ANOVA

There are two factors, say R and C. We observe random variables X_{ij} which are mutually independent. They have possible values of R_i, C_j. We observe a process acting on certain combination of R_i and C_j.We can extend this idea to $n-$way ANOVA.

For example, R are treatments and C are different patients.

Now, suppose $X_{ij} \sim N(\xi_{ij}, \sigma^2)$, $i = 1, \cdots, n$ and $j = 1, \cdots, s$. And σ^2 is unkown.

So $\xi_{ij} = \mathbb{E}X_{ij}$ and $\xi_{ij} = \mu + \alpha_i + \beta_j$, where μ is the general parameter, α_i is the corresponding effect of R_i and β_j is the corresponding effect of C_j.

When saying the additive combination $\alpha_i + \beta_j$, we assume there is no interaction between these two factors, or they are independent.

Note that a more general form is $\xi_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ which allows interaction between two factors.

WLOG, assume that $\sum_{i=1}^r \alpha_i = 0$ and $\sum_{j=1}^s \beta_j = 0$. We can do this we can always set $\mu' = \mu + \sum_{i=1}^r \alpha_i + \sum_{j=1}^s \beta_j$, $\alpha_i' = \alpha_i - \sum_{i=1}^r \alpha_i$ and $\beta_j' = \beta_j - \sum_{j=1}^s \beta_j$.

In total, we have $r + s - 1 - 1 + 1 + 1 = r + s$ parameters for α, β, μ and σ.

And $X_{ij} \overset{d}{=} \mu + \alpha_i + \beta_j + \varepsilon_{ij}\sigma$ and $\varepsilon_{ij} \sim N(0,1)$. This is similar to a linear regression.

A typical question in ANOVA is to test whether $\alpha_1 = \cdots = \alpha_r = 0$ or not? Similarly, we can also test whether $\beta_1 = \cdots = \beta_s = 0$ or not?

We can use LSE to estimate α_i, β_j and μ.

We need to minimize $S = \sum_{i=1}^r \sum_{j=1}^s (X_{ij} - \xi_{ij})^2$. Introduce notations $\bar{X}_{\cdot\cdot} = \frac{1}{rs}\sum_{i,j} X_{ij}$, $\bar{X}_{i\cdot} = \frac{1}{s}\sum_j X_{ij}$ and $\bar{X}_{\cdot j} = \frac{1}{r}\sum_j X_{ij}$, so $\sum_{i=1}^r \bar{X}_{i\cdot} = r\bar{X}_{\cdot\cdot}$ and $\sum_{j=1}^s \bar{X}_{\cdot j} = s\bar{X}_{\cdot\cdot}$

Then $S = \sum_{i,j}(X_{ij} - \mu - \alpha_i - \beta_j)^2 = \sum_{i,j}(X_{ij} - \mu - \alpha_i - \beta_j)^2$
$= \sum_{i,j}[(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot}) + (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot} - \alpha_i) + (\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot} - \beta_j) + (\bar{X}_{\cdot\cdot} - \mu)]^2$
$= \sum_{i,j}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot})^2 + s\sum_{i=1}^r(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot} - \alpha_i)^2 + r\sum_{j=1}^s(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot} - \beta_j)^2 + rs(\bar{X}_{\cdot\cdot} - \mu)^2$ since all cross products are 0.

Then S attains its minimum at $\hat{\alpha}_i = \bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}, \hat{\beta}_j = \bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot}$ and $\hat{\mu} = \bar{X}_{\cdot\cdot}$. We have $\hat{\xi}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$. And $S_{\min} = \sum_{i,j}(X_{ij} - \hat{\xi}_{ij})^2 = \sum_{i,j}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot})^2$ where $\hat{\mu} = \frac{1}{sr}\sum_{i,j} X_{ij} \sim N(\mu, \frac{\sigma^2}{sr})$ and $\xi_{\cdot\cdot} = \frac{1}{sr}\sum_{i,j}\xi_{ij} = \mu$.

For $\hat{\alpha}_i$, we have $\bar{X}_{i\cdot} = \frac{1}{s}\sum_{j=1}^s X_{ij}$, $\mathbb{V}\mathrm{ar}(\bar{X}_{i\cdot}) = \frac{\sigma^2}{s}$ and $\mathbb{E}\bar{X}_{i\cdot} = \mu = \mathbb{E}\bar{X}_{\cdot j}$.

Also, $\mathbb{V}\text{ar}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}) = \mathbb{V}\text{ar}(\bar{X}_{i\cdot} - \frac{1}{r}\sum_{k=1}^{n}\bar{X}_{k\cdot}) = \mathbb{V}\text{ar}(\bar{X}_{i\cdot}(1-\frac{1}{r}) - \frac{1}{r}\sum_{k\neq i}\bar{X}_{k\cdot}) = (\frac{r-1}{r})^2\frac{\sigma^2}{s} + \frac{r-1}{r^2}\frac{\sigma^2}{s} = \sigma^2\frac{r-1}{rs}$. So $\hat{\alpha}_i \sim N(0, \sigma^2\frac{r-1}{rs})$ and similarly $\hat{\beta}_j \sim N(0, \sigma^2\frac{s-1}{rs})$.

In order to construct a confidence set independent of σ^2, we use the student's theorem for $X_i \sim N(\mu, \sigma^2)$, which states that $\frac{\bar{X}-\mu}{\sqrt{S/n}} \sim t_{n-1}$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ is independent to $\bar{X}$.

In our case, this reduces to

$$\frac{(\hat{\mu}-\mu)\sqrt{rs(r-1)(s-1)}}{\sqrt{S_{\min}}} \sim t_{(r-1)(s-1)}$$

This property follows for example, from the result for LSE in linear Gaussian regression, $\frac{\hat{\theta}_j-\theta_j}{\hat{\sigma}\sigma_{jj}}\sqrt{\frac{n-p}{n}} \sim t_{n-p}$. The model is $Y_i = X_i^T\theta + \underbrace{\varepsilon_i}_{\sim N(0,\sigma^2)}$, $\theta \in \mathbb{R}^p$. And $q^2 = \frac{1}{n}S_{\min} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\theta)^2$, $\sigma_{jj}^2 = \mathbb{V}\text{ar}(\frac{\hat{\theta}_j}{\sigma^2}) = (X^TX)^{-1}_{jj}$.

And

$$\frac{\hat{\alpha}_i\sqrt{rs(s-1)}}{\sqrt{S_{\min}}} \sim t_{(r-1)(s-1)}$$

$$\frac{\hat{\beta}_j\sqrt{rs(r-1)}}{\sqrt{S_{\min}}} \sim t_{(r-1)(s-1)}$$

Also, $\frac{S_{\min}}{(r-1)(s-1)}$ is an unbiased estimator of S^2, and $\frac{S_{\min}}{\sigma^2} \sim \chi^2_{(r-1)(s-1)}$.

Now, consider the following hypothesis. $H_0 : \alpha_1 = \cdots = \alpha_r = 0$ (a linear constraint). And $\sum_i \alpha_i = 0, \sum_j \beta_j = 0$ is a linear space in $\mathbb{R}^{r+s-2}$.

$S = \sum_{i,j}(X_{ij} - \xi_{ij})^2$ and

$$S_T = \min_{\alpha_1=\cdots=\alpha_r=0} S = S_{\min} + s\sum_{i=1}^{r}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 \geq S_{\min}$$

Let $F = \frac{(S_T - S_{\min})/(r-1)}{S_{\min}/[(r-1)(s-1)]} \sim F_{r-1,(r-1)(s-1)}$ under H_0. Note that $S_T - S_{\min} \sim \sigma^2\chi^2_{r-1}$ and $S_T \sim \sigma^2\chi^2_{(r-1)(s-1)}$ are independent by Cochran theorem. Also $S_T - S_{\min} \sim \chi^2$ if H_0 is true and it will become larger than $s\sum_{i=1}^{r}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$ if H_0 is false.

We reject H_0 at significance level α if $F \geq$ upper quantile of the $F_{(r-1)(s-1)}$ distribution.

Another hypothesis may be $\alpha_1 = \cdots = \alpha_r = \beta_1 = \cdots = \beta_s = 0$ and $F = \frac{(r-1)(s-1)}{r+s-2}\frac{s\sum_{i=1}^{r}(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})^2 + r\sum_{j=1}^{s}(\bar{X}_{\cdot j}-\bar{X}_{\cdot\cdot})^2}{S_{\min}} \sim F_{r+s-2,(r-1)(s-1)}$ under H_0.

5 High-dimensional Linear Models

Let $Y = X\beta + \xi$, $y \in \mathbb{R}^n, \beta \in \mathbb{R}^N$ where X is a $n \times N$ matrix and $\xi \sim N(0, \sigma^2 I_n)$. Let $x_j, j = 1, \cdots, N$ be the columns of X, then $Y = \sum_{j=1}^{N}\beta_j x_j + \xi$.

Introduce $L = \text{linear span}(x_1, \cdots, x_N) \subset \mathbb{R}^n$. Also $X_1, \cdots, X_n$ be the rows of X. And $Y_i = \langle X_i, \beta \rangle + \xi_i, i = 1, \cdots, n$. This is called the n noisy-linear measurements of β and ξ are iid $N(0, \sigma^2 I_n)$.

In image processing, we can often use some bases such as Fourier bases or wavelet bases and there are many coefficients. Therefore, the idea of compressed sensing is introduced.

The least-square estimator is $\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^N} ||Y - X\beta||^2$. We know that $X\hat{\beta} = P_L Y$. The error of this estimator is $\mathbb{E}||X\hat{\beta} - X\beta||^2 = \sigma^2 \dim(L)$.

If the X columns are linear independent, and the design matrix is full-rank rank$(X) = N$, or $n \geq N$, then β is identifiable.

And by proposition 4.1, $\mathbb{E}||\hat{\beta} - \beta||^2 = \sigma^2 tr((X^T X)^{-1}) \leq \sigma^2 N \lambda_{\max}(X^T X)^{-1} = \frac{\sigma^2 N}{\lambda_{\min}(X^T X)}$ where $X^T X_{ij} = \langle x_i, x_j \rangle$ is strictly positive definite. In particular, if X is an orthogonal design, then $X^T X = I_N \Leftrightarrow x_1, \cdots, x_N$ will be orthonormal systems in $\mathbb{R}^n$, and $\lambda_{\min}(X^T X) = 1$, so $\mathbb{E}||\hat{\beta} - \beta||^2 = \sigma^2 N$. We are in trouble if $n < N$ or if we are unable to collect that many samples.

Definition 5.1. degree of sparsity

Let $J_\beta = \text{supp}(\beta) = \{j = 1, \cdots, N, \beta_j \neq 0\}, \beta \in \mathbb{R}^N$ and $d(\beta) = \text{card}(J_\beta) = \sum_{j=1}^N \mathbb{I}(\beta_j \neq 0)$ is called the **degree of sparsity** of vector β. If $d(\beta) << N$, we say that β is sparse.

Then the problem is to $\min_{\beta \in \mathbb{R}^N} d(\beta)$ subject to $X\beta = Y$. We want to show that we can transform this non-convex problem to convex, $\min_{\beta \in \mathbb{R}^N} ||\beta||_1$ subject to $X\beta = Y$.

A natural question is whether there exists an estimator $\hat{\beta}$ such that $\mathbb{E}||\hat{\beta} - \beta||^2 \leq \sigma^2 d(\beta)$.

A typical penalized least square estimators is $\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^N} ||Y - X\beta||^2 + c\sigma^2 d(\beta)$ where $c\sigma^2 d(\beta)$ is called the **complexity of penalty** for lack of sparsity.

It can also be used to do variable selection. To solve the above problem, we choose a subset $I \subset \{1, \cdots, N\}$ and we have 2^N choices. We first solve $\hat{\beta}_I = \text{argmin}_{\beta \in \mathbb{R}^N,\ \text{supp}(\beta) = I} ||Y - X\beta||^2$. Then minimize $||Y - X\hat{\beta}_I||^2 + c\sigma^2 d(I)$ over all possible subsets I. But this is rather computationally intensive.

Definition 5.2. LASSO estimator

The LASSO estimator is defined as $\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^N} ||Y - X\beta||^2 + c\sigma^2 ||\beta||_{l_1}$ where $||\beta||_{l_1} = \sum_{i=1}^N |\beta_j|$.

We first look at noiseless cases: or **sparse recovery problem**.

Again we have a $n \times N$ design matrix X and we want to solve $X\beta = Y$ where $Y \in \mathbb{R}^n, \beta \in \mathbb{R}^N$ and $n << N$. In another way, $\sum_{j=1}^N \beta_j x_j = Y$ so we have N unkown variables. The solution will have dimensions of $N - n$.

Now $Y_j = \langle X_j, \beta \rangle, j = 1, \cdots, n$ or we have n noiseless linear measurements of β and define $M = \{u \in \mathbb{R}^N, Xu = Y\}$ which is an affine subspace of $\mathbb{R}^N$ of all solutions of the linear system. We want to minimize $d(u)$ subject $u \in M$. Equivalently, find the sparsest vector in M.

We first look at the problem to minimize $||u||_{l_1}$ over all $u \in M$. This is a **linear programming** since it can be rephrased as $\min \sum_{i=1}^N c_i$ such that $c_j \geq 0, -c_i \leq u_j \leq c_j$ and $Xu = Y$. It is a convex problem, what's better, it's a linear programming problem. Then we can use some theorems to convert the problem to the orginal $d(u)$ minimization.

Under **Restricted Isometry Property**, $\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^N} ||u||_{l_1} \implies \hat{\beta} = \beta$ provided β is sufficiently sparse.

Definition 5.3. Let $J \subset \{1, \cdots, N\}$ and define the cone $C_J = \{u \in \mathbb{R}^N : \sum_{j \notin J} |u_j| \leq \sum_{j \in J} |u_j|\}$. For example $\{(x,y) : |x| \leq |y|\}$ is a cone in $\mathbb{R}^2$.

Definition 5.4. Let J be a subset and X be a $n \times N$ matrix, define $\gamma(J,X) = \inf\{C > 0 : \sum_{j \in J} u_j^2 \leq C^2 ||Xu||^2, u \in C_J\}$. Or we are tyring to bound $||Xu||$ away from 0.

Theorem 5.1. *Suppose we have $Y = X\beta$ and $\gamma(J_\beta, X) < \infty$ and let $\hat{\beta} = \text{argmin}_{Xu=Y} ||u||_{l_1}$. The claim is $\hat{\beta} = \beta$. We know that l_1 norm is convex so $\hat{\beta}$ is unique, and the theorem tells that β and $\hat{\beta}$ are both* ***unique***.

Proof

Let $\hat{u} = \hat{\beta} - \beta$. Note that $X\hat{\beta} = Y, X\beta = Y \implies X(\hat{\beta} - \beta) = 0 \implies X\hat{u} = 0 \implies \hat{u} \in Ker(X)$.

The second fact is $\hat{u} \in C_{J_\beta}$. To check this, note that by definition,

$||\hat{\beta}||_{l_1} \leq ||\beta||_{l_1} \Leftrightarrow \sum_{j=1}^N |\hat{\beta}_j| \leq \sum_{j=1}^N |\beta_j| = \sum_{j \in J_\beta} |\beta_j| \implies \sum_{j \notin J_\beta} |\hat{\beta}_j| \leq \sum_{j \in J_\beta} (|\beta_j| - |\hat{\beta}_j|) \leq \sum_{j \in J_\beta} |\hat{\beta}_j - \beta_j|$

Since for $j \notin J_\beta, \beta_j = 0$, we have $\sum_{j \notin J_\beta} |\underbrace{\hat{\beta}_j - \beta_j}_{\hat{u}_j}| \leq \sum_{j \in J_\beta} |\underbrace{\hat{\beta}_j - \beta_j}_{\hat{u}_j}| \implies \hat{u} \in C_\beta$.

Since $\gamma(J_\beta, X) < \infty$, we have $(\sum_{j \in J_\beta} (u_j)^2)^{\frac{1}{2}} \leq \gamma(J_\beta, X) ||\underbrace{X\hat{u}}_{0}|| = 0$ by definition.

It follows that $\sum_{j \in J_\beta} (\hat{u}_j)^2 = 0 \implies |\hat{u}_j| = 0, j \in J_\beta, \sum_{j \in J_\beta} |\hat{u}_j| = 0$.

Also, $\hat{u} \in C_{J_\beta} \implies \sum_{j \notin J_\beta} |\hat{u}_j| \leq \sum_{j \in J_\beta} |\hat{u}_j| \implies |\hat{u}_j| = 0, j \notin J_\beta \implies \hat{\beta} = \beta$. □

Definition 5.5. isometry property

Recall that X is orthogonal design $\Leftrightarrow X^T X = I_N \Leftrightarrow x_1, \cdots, x_n$ are orthonormal. And $||Xu||^2 = ||u_1 x_1 + \cdots + u_N x_N||^2 = ||u||^2$ by Pythagorean theorem. So $||Xu|| = ||u||, u \in \mathbb{R}^N$ and this is called **isometry property**.

Definition 5.6. Restricted isometry constant

Restricted isometry constant, introduced by Emmanuel Candes, Justin Romberg and Terence Tao, is defined as $\delta_d(X) = \inf_{\delta > 0}\{u \in \mathbb{R}^N, d(u) \leq d, 1 - \delta \leq \frac{||Xu||^2}{||u||^2} \leq 1 + \delta\}$. It's clear that $\delta(d)$ is non-decreasing with respect to d.

Proposition 5.1. If $L_1, L_2 \subset V$, define $p = \sup_{x\in L_1, y\in L_2, x,y\neq 0} \frac{|\langle x,y\rangle|}{||x||\cdot||y||}$. Then for $\forall x \in L_2$, $||P_{L_1}x|| \leq p||x||$.

Proposition 5.2. For $u, v \in \mathbb{R}^N$ with $d(u) \leq d, d(v) \leq d$ such that $\mathrm{supp}(u) \cap \mathrm{supp}(v) = \emptyset$, then $|\cos(Xu, Xv)| = \frac{|\langle Xu,Xv\rangle|}{||Xu||\cdot||Xv||} \leq c\delta_{2d}(X)$. Or the angles between Xu and Xv are close to 90 degress.

Proposition 5.3. Suppose $\delta_{3d}(X) \leq c$ where $c > 0$ is a small numerical constant. Then for any β with $d(\beta) \leq d$, $\gamma(J_\beta, X) < \infty$. Or we can recover any vector β with $d(\beta) \leq d$ and $Y = X\beta$ using $\beta = \hat{\beta} = \mathrm{argmin}_{Xu=Y} ||u||_{l_1}$.

Proof

Recall $C_J = \{u \in \mathbb{R}^N, \sum_{j\notin J} |u_j| \leq \sum_{j\in J} |u_j|\}$. Suppose now $\mathrm{card}(J) = d << N$. Then consider representation of vectors $u \in C_J$ as sum of $d-$sparse vectors.

First, set $J_0 = J$, for any $u \in C_J$, arrange $|u_j|$ for $j \in \{1, \cdots, N\} \setminus J_0$ in non-increasing order.

And J_1 be the set of d largest coordinates in $j \in \{1, \cdots, N\} \setminus J_0$.

And J_2 be the set of d next coordinates in $j \in \{1, \cdots, N\} \setminus (J_0 \cup J_1)$.

Keep doing this until running out of coordinates.

Now, define $u^{(0)} = (u_j, j \in J_0)$, $u^{(1)} = (u_j, j \in J_1)$ and $u^{(k)} = (u_j, j \in J_k)$. Note that $u = u^{(0)} + u^{(1)} + \cdots$ and $d(u^{(i)}) \leq d$ for all i.

<u>Claim:</u> For $u \in C_J$, $\sum_{k\geq 2} ||u^{(k)}|| \leq ||u^{(0)}||$.

Proof

For any $k \geq 1$ and $j \in J_{k+1}$, $|u_j| \leq \min_{i\in J_k} |u_i| \leq \frac{1}{d} \sum_{i\in J_k} |u_i|$ by our construction of J_k.

And $\sum_{j\in J_{k+1}} |u_j|^2 \leq d\frac{1}{d^2}(\sum_{i\in J_k} |u_i|)^2 = \frac{1}{d}(\sum_{i\in J_k} |u_i|)^2$. So $||u^{(k+1)}|| = (\sum_{j\in J_{k+1}} |u_j|^2)^{1/2} \leq \frac{1}{\sqrt{d}} \sum_{i\in J_k} |u_i|$.

Hence $\sum_{k\geq 2} ||u^{(k)}|| = \sum_{k\geq 1} ||u^{(k+1)}|| \leq \frac{1}{\sqrt{d}} \sum_{k\geq 1} \sum_{i\in J_k} |u_i| = \frac{1}{\sqrt{d}} \sum_{i\notin J} |u_i| = \frac{1}{\sqrt{d}} \sum_{i\in J} |u_i| \cdot 1 \leq \underbrace{\frac{1}{\sqrt{d}} (\sum_{i\in J} |u_j|^2)^{1/2} \sqrt{d}}_{\text{Cauchy Schwarz}} =$

$||u^{(0)}||$.

□

Proposition 5.4. For any $u \in C_J$, $u = u^{(0)} + u^{(1)} + \cdots = ||u^{(0)}||\underbrace{\frac{u^{(0)}}{||u^{(0)}||}}_{v^{(0)}} + ||u^{(1)}||\underbrace{\frac{u^{(1)}}{||u^{(1)}||}}_{v^{(1)}} + \ldots$. Note that $||v^{(i)}||= 1$ and $d(v^{(i)}) \leq d$ and for any $u \in C_J \cap \{||u||= 1\}$, $u = \sum_{k\geq 0}||u^{(k)}||v^{(k)}$ and $\sum_{k\geq 0}||u^{(k)}||= ||u^{(0)}||+||u^{(1)}||+\underbrace{\sum_{k\geq 2}||u^{(k)}||}_{\leq ||u^{(0)}||} \leq 3$.

Therefore, we have the following corollary:

Corollary 5.2. *For* $C_J \cap \{||u||\leq 1\} \subset 3\mathrm{conv}(\{v : ||v||= 1, d(v) \leq d\}$

We have $\gamma(J,X) = \inf\{C > 0 : \sum_{j\in J}(u_j)^2 \leq C^2||Xu||^2, u \in C_J\}$.

Suppose $\mathrm{card}(J) = d$, then $||Xu||= ||\sum_j u_j x_j||$ where x_j are columns of the design matrix X. For any $I \subset \{1, \cdots, N\}$, $L_I =$ linear span$(x_j, j \in I)$ and now let's introduce the projection operator $P_I = P_{L_I}$.

For any $u \in C_J$,

$||Xu||\geq ||P_{J_0\cup J_1}Xu||= ||P_{J_0\cup J_1}\sum_{k\geq 0}Xu^{(k)}||$
$= ||\underbrace{P_{J_0\cup J_1}(Xu^{(0)} + Xu^{(1)})}_{Xu^{(0)}+Xu^{(1)}} + P_{J_0\cup J_1}\sum_{k\geq 2}Xu^{(k)}||$
$\geq ||(Xu^{(0)} + Xu^{(1)})||-||P_{J_0\cup J_1}\sum_{k\geq 2}Xu^{(k)}||$
$\geq ||(Xu^{(0)} + Xu^{(1)})||-\sum_{k\geq 2}||P_{J_0\cup J_1}Xu^{(k)}||$
$\geq ||(Xu^{(0)} + Xu^{(1)})||-c'\delta_{3d}\sum_{k\geq 2}||Xu^{(k)}||$ (Take $u = u^{(0)} + u^{(1)}, v = u^{(k)}, k \geq 2$, then $\mathrm{supp}(u)\cap \mathrm{supp}(v) = \emptyset$, and by proposition 5.2, $\frac{|\langle Xu,Xv\rangle|}{||Xu||\cdot||Xv||} \leq c'\delta_{2d}(X) \leq c'\delta_{3d}(X)$. Take $\{Xu, u = u^{(0)} + u^{(1)}\} = L_0 + L_1$ and $\{Xv, v = u^{(k)}\} = L_k$. Then by proposition 5.1, we have $\sup_{x\in(L_0+L_1),y\in L_k}\frac{\langle x,y\rangle}{||x||\cdot||y||} = \sup_{Xu,Xv}\frac{\langle Xu,Xv\rangle}{||Xu||\cdot||Xv||} \leq c'\delta_{3d}(X) \implies ||P_{J_0\cup J_1}Xu^{(k)}||\leq c'\delta_{3d}(X)||Xu^{(k)}||$ which holds for every $k \geq 2$, since $Xu^{(k)} \in L_k$.)

$\geq ||(Xu^{(0)} + Xu^{(1)})||-\sum_{k\geq 2}c'\delta_{3d}\sqrt{1+\delta_d}||u^{(k)}||$ because $||Xu^{(k)}||\leq \sqrt{1+\delta_d}||u^{(k)}||$ since $d(u^{(k)}) \leq d$ and by the definition of restricted isometry constant.

$\geq ||(Xu^{(0)} + Xu^{(1)})||-c'\delta_{3d}\sqrt{1+\delta_d}||u^{(0)}||$
$\geq ||Xu^{(0)} + Xu^{(1)}||-c'\delta_{3d}\sqrt{1+\delta_d}||u^{(0)} + u^{(1)}||$

$\geq ||Xu^{(0)} + Xu^{(1)}||-\frac{c'\delta_{3d}\sqrt{1+\delta_d}}{\sqrt{1-\delta_{2d}}}||Xu^{(0)} + Xu^{(1)}||= ||Xu^{(0)} + Xu^{(1)}||(1-\frac{c'\delta_{3d}\sqrt{1+\delta_d}}{\sqrt{1-\delta_{2d}}})$ since $||u^{(0)} + u^{(1)}||\leq \frac{||Xu^{(0)}+Xu^{(1)}||}{\sqrt{1-\delta_{2d}}}$

$\geq (\sqrt{1-\delta_{2d}})(1-\frac{c'\delta_{3d}\sqrt{1+\delta_d}}{\sqrt{1-\delta_{2d}}})||u^{(0)} + u^{(1)}||\geq (1-\frac{c'\delta_{3d}(1+\delta_d)}{1-\delta_{2d}})(1-\delta_{2d})||u^{(0)}||\geq (\sqrt{1-\delta_{2d}}-c'\delta_{3d}\sqrt{1+\delta_d})(\sum_{j\in J}|u_j|^2)^{1/2}$

$\implies \sum_{j\in J}|u_j|^2\leq (\frac{1}{\sqrt{1-\delta_{2d}}-c'\delta_{3d}\sqrt{1+\delta_d}})^2||Xu||^2 \implies \gamma(J,X) \leq \frac{1}{\sqrt{1-\delta_{2d}}-c'\delta_{3d}\sqrt{1+\delta_d}} < \infty$

Also, δ_{3d} should be small so that every term in the proof is positive. □

Proposition 5.5. Suppose $\delta_{3d}(X) \leq c$ where c is a small constant. Let $Y = X\beta, d(\beta) \leq d$, then the equation β is the only d-sparse solution of the equation $Xu = Y$. Moreover, $\beta = \text{argmin}_{Xu=Y}||u||_{l_1}$. Unfortunately, it's not easy to construct deterministic X such that $\delta_{3d}(X) \leq c$ but we can construct stochastic matrices satisfying $\delta_{3d}(X) \leq c$ with high probability. Details can be seen at theorem 5.5.

Restricted Isometry Property means $1-\delta \leq \frac{||Xu||^2}{||u||^2} \leq 1+\delta$ for any $u, d(u) \leq d$.

Let's assume we are in some subspace and ignore the condition $d(u) \leq d$ for now.

Note that $||Xu||^2 = \langle Xu, Xu\rangle = \langle X^T Xu, u\rangle$ where $X^T X$ is a symmetric matrix and let's call its eigenvalues $\lambda_1(X^T X) \leq \cdots \leq \lambda_N(X^T X) \geq 0$. We basically want $\sup_{u\neq 0} \frac{||Xu||^2}{||u||^2} = \sup_{||u||=1} \frac{||Xu||^2}{||u||^2} = \sup_{||u||=1}\langle X^T Xu, u\rangle = \lambda_1(X^T X)$ and $\inf_{u\neq 0} \frac{||Xu||^2}{||u||^2} = \inf_{||u||=1} \frac{||Xu||^2}{||u||^2} = \inf_{||u||=1}\langle X^T Xu, u\rangle = \lambda_N(X^T X)$.

Definition 5.7. $\sigma_j(X) = \sqrt{\lambda_j(X^T X)}$ is called the j-th singular value of X.

So $1-\delta \leq \frac{||Xu||^2}{||u||^2} \leq 1+\delta \Leftrightarrow \sigma_{\max}(X) \leq \sqrt{1+\delta}, \sigma_{\min}(X) \geq \sqrt{1-\delta}$.

We now look at a theorem about bounds on singular values of X.

Definition 5.8. Operator norm

Suppose A is a symmetric matrix, then the **operator norm** or **spectral norm** of A is defined as $||A|| = \sup_{||u||\leq 1}||Au|| = \sup_{||u||\leq 1}\langle Au, u\rangle = \max_{1\leq j\leq N}|\lambda_j(X)|$. where $\lambda_j(A)$ are the eigenvalues of A.

Proposition 5.6. $|\sigma_{\max}(X)-1| \leq \min(||X^T X - I||, \sqrt{||X^T X - I||})$ and $|\sigma_{\min}(X)-1| \leq \min(||X^T X - I||, \sqrt{||X^T X - I||})$. The key point is to note that for any $a \geq 0$, $|a-1| \leq \min(|a^2-1|, \sqrt{|a^2-1|})$.

Theorem 5.3. *Bernstein inequality*

Let $\xi_1, \cdots, \xi_n$ be the independent normal variables with $N(0, \sigma^2)$, then for $t > 0$, with probability $1-e^{-t}$, the following will be true: $|\frac{1}{n}\sum_{j=1}^n(\xi_j^2 - \mathbb{E}\xi_j^2)| \lesssim \sigma^2(\sqrt{\frac{t}{n}} \vee \frac{t}{n})$ *where* $a \vee b = \max(a,b)$.

Theorem 5.4. *Let X be a $n \times N$ matrix and* $X = \begin{bmatrix} \frac{X_1}{\sqrt{n}} \\ \cdots \\ \frac{X_n}{\sqrt{n}} \end{bmatrix}$ *where X_i are iid $N(0, I_N)$ with entries X_{ij} iid $N(0, \frac{1}{\sqrt{n}})$. For* *any $t > 1$, the following bounds hold with probability at least $1 - e^{-t}$:*

$|\sigma_{\max}(X) - 1| \lessapprox \sqrt{\frac{N}{n}} + \sqrt{\frac{t}{n}}$ *and* $|\sigma_{\min}(X) - 1| \lessapprox \sqrt{\frac{N}{n}} + \sqrt{\frac{t}{n}}$. *where* $\lessapprox$ *means the less than is up to a constant.*

Proof

$\|X^TX-I\| = \sup_{\|u\|\le 1}\langle (X^TX-I)u,u\rangle$. Note that $\langle (X^TX-I)u,u\rangle = \langle X^TXu,u\rangle - \|u\|^2 = \langle Xu,Xu\rangle - \|u\|^2 = \|Xu\|^2-\|u\|^2 = \sum_{j=1}^n \langle Xu,e_j\rangle^2 - \|u\|^2 = \sum_{j=1}^n \langle u, \underbrace{X^Te_j}_{j \text{ th row of X}}\rangle^2 - \|u\|^2 = \sum_{j=1}^n \langle \frac{X_j}{\sqrt{n}},u\rangle^2 - \|u\|^2 = \frac{1}{n}\sum_{j=1}^n \langle X_j,u\rangle^2 - \|u\|^2 = \frac{1}{n}\sum_{j=1}^n(\langle X_j,u\rangle^2 - \mathbb{E}\langle X_j,u\rangle^2)$ since $\mathbb{E}\langle X_j,u\rangle^2 = \|u\|^2$ and if we choose $e_1,\cdots,e_n$ be the canonical orthonormal bases in $\mathbb{R}^n$.

By homework, $|\sigma_{\max}(X)-1| \le \|X^TX-I\|$ and $|\sigma_{\min}(X)-1| \le \|X^TX-I\|$ and here we use the operator norm.

Discretization: By homework, there exists a subset $M \subset \{u\in\mathbb{R}^N : \|u\|\le 1\}$ such that $\text{card}(M) \le 9^N$ and for any u such that $\|u\|\le 1$ there exits $u' \in M : \|u-u'\| \le \frac{1}{4}$, or M is a $\frac{1}{4}$-net for $\{u : \|u\|\le 1\}$ of card $\le 9^N$.

<u>Claim:</u> $\|X^TX-I\| \le 2\max_{u\in M}|\langle (X^TX-I)u,u\rangle|$.

Proof

For any u such that $\|u\|\le 1$ there exists $u'\in M$ such that $\|u-u'\|\le \frac{1}{4}$.

Let's consider the cost to replace u to u', $|\langle (X^TX-I)u,u\rangle - \langle (X^TX-I)u',u'\rangle|$
$\le |\langle (X^TX-I)u,u\rangle - \langle (X^TX-I)u',u\rangle| + |\langle (X^TX-I)u',u\rangle - \langle (X^TX-I)u',u'\rangle|$
$= |\langle (X^TX-I)(u-u'),u\rangle| + |\langle (X^TX-I)u',u-u'\rangle|$
$\le \|X^TX-I\|\cdot \underbrace{\|u-u'\|}_{\le\frac{1}{4}}\cdot\underbrace{\|u\|}_{\le 1} + \|X^TX-I\|\cdot\underbrace{\|u'\|}_{\le 1}\cdot\underbrace{\|u-u'\|}_{\le\frac{1}{4}} \le \frac{1}{2}\|X^TX-I\|$.

Now $\|X^TX-I\| = \sup_{\|u\|\le 1}|\langle (X^TX-I)u,u\rangle| \le \max_{u'\in M}|\langle (X^TX-I)u',u'\rangle| + \frac{1}{2}\|X^TX-I\|$ by the previous parts.

$\implies \|X^TX-I\| \le 2\max_{u\in M}|\langle (X^TX-I)u,u\rangle|$

□

Recall that $\langle (X^TX-I)u,u\rangle = \frac{1}{n}\sum_{j=1}^n(\langle X_j,u\rangle^2 - \mathbb{E}\langle X_j,u\rangle^2)$. Since $\langle X_j,u\rangle$ are independent $N(0,\underbrace{\|u\|^2}_{\le 1})$, by Bernstein inequality, with probability $\ge 1-e^{-t}$, $|\frac{1}{n}\sum_{j=1}^n\langle X_j,u\rangle^2 - \|u\|^2| \lessapprox \sqrt{\frac{t}{n}} \vee \frac{t}{n}$ for each fixed u.

By using the probability union bound, for $\forall u\in M$, $|\langle (X^TX-I)u,u\rangle| \lessapprox \sqrt{\frac{t}{n}}\vee\frac{t}{n}$ with probability $\ge 1-\text{card}(M)e^{-t}$ where $\text{card}(M)$ is the number of points in M. This is true since for one particular u, the probability of the event that $|\langle (X^TX-I)u,u\rangle| \lessapprox \sqrt{\frac{t}{n}}\vee\frac{t}{n}$ doesn't hold is e^{-t} and for arbitrary u, the probability this doesn't hold is less or equal to $\text{card}(M)e^{-t}$.

Then with probability $\ge 1-\text{card}(M)e^{-t}$, $\|X^TX-I\| \le 2\max_{u\in M}|\langle (X^TX-I)u,u\rangle| \lessapprox \sqrt{\frac{t}{n}}\vee\frac{t}{n}$.

Let's now replace t with $t+\log\underbrace{\mathrm{card}(M)}_{\leq 9^N}$ or even more, to $t+N\log 9$, then with probability $\geq 1-\mathrm{card}(M)e^{-t-\log\mathrm{card}(M)} = 1-e^{-t}$, we have $||X^TX-I|| \lessapprox \sqrt{\frac{t+N}{n}} \vee \frac{t+N}{n}$.

From homework, $|\sigma_{\max}(X)-1| \leq ||X^TX-I|| \wedge ||X^TX-I||^{1/2}$ and $|\sigma_{\min}(X)-1| \leq ||X^TX-I|| \wedge ||X^TX-I||^{1/2}$.

We know with probability $\geq 1-e^{-t}$, $||X^TX-I|| \lessapprox \sqrt{\frac{t+N}{n}} \vee \frac{t+N}{n}$

$\implies |\sigma_{\max}(X)-1| \leq ||X^TX-I|| \wedge ||X^TX-I||^{1/2} \lessapprox (\sqrt{\frac{t+N}{n}} \vee \frac{t+N}{n}) \wedge (\sqrt{\frac{t+N}{n}} \vee \frac{t+N}{n})^{1/2} = \sqrt{\frac{t+N}{n}}$.

We then get with probability $\geq 1-e^{-t}$, $|\sigma_{\max}(X)-1| \lessapprox \sqrt{\frac{t+N}{n}}$ and $|\sigma_{\min}(X)-1| \lessapprox \sqrt{\frac{t+N}{n}}$. $\square$

Theorem 5.5. *Let X be a $n\times N$ matrix and* $X = \begin{bmatrix} \frac{X_1}{\sqrt{n}} \\ \cdots \\ \frac{X_n}{\sqrt{n}} \end{bmatrix}$ *where X_i are iid $N(0,I_N)$ with entries X_{ij} iid $N(0,\frac{1}{\sqrt{n}})$.*

Suppose d satisfies $\sqrt{\frac{d\log N/d}{n}} \leq c'$ (small constant). Then with high probability (to be specified), $\delta_d(X) \leq c$. More precisely, we can say that for any c, there exists a c' such that the inequality holds.

Proof

Recall that $\delta_d(X) = \inf_{\delta>0}\{u\in\mathbb{R}^N, d(u)\leq d, 1-\delta \leq \frac{||Xu||^2}{||u||^2} \leq 1+\delta\}$ and suppose $\mathrm{supp}(u)\subset I$ and $\mathrm{card}(I)=d$.

Let $X_I = (x_j : j\in I)$, or we pick columns belong to I from X.

Then $1-\delta \leq \frac{||Xu||^2}{||u||^2} \leq 1+\delta \Leftrightarrow 1-\delta \leq \frac{\langle X^TXu,u\rangle}{||u||^2} \leq 1+\delta \Leftrightarrow$ eigenvalues of $X_I^TX_I \in (1-\delta,1+\delta)$

$\Leftrightarrow$ singular values of $X_I \in (\sqrt{1-\delta},\sqrt{1+\delta})$

$\Leftrightarrow \sqrt{1-\delta} \leq \sigma_{\min}(X_I) \leq \sigma_{\max}(X_I) \leq \sqrt{1+\delta}$

From the previous bounds $|\sigma_{\max}(X)-1| \lessapprox \sqrt{\frac{t+N}{n}}$ and $|\sigma_{\min}(X)-1| \lessapprox \sqrt{\frac{t+N}{n}}$, for any $I\subset\{1,\cdots,N\}$, $\mathrm{card}(I)\leq d$, with probability $\geq 1-e^{-t}$, $|\sigma_{\max}(X_I)-1| \lessapprox \sqrt{\frac{t+d}{n}}$ and $|\sigma_{\min}(X_I)-1| \lessapprox \sqrt{\frac{t+d}{n}}$.

Let $J_d = \{I\subset\{1,\cdots,N\} : \mathrm{card}(I)\leq d\}$ and $\mathrm{card}(J_d) = \sum_{k=1}^d \binom{n}{k} = \binom{n}{\leq d} \leq (\frac{eN}{d})^d$.

By the union bound, with probability $\geq 1-\mathrm{card}(J_d)e^{-t}$, $\max_{I\in J_d}|\sigma_{\max}-1| \lessapprox \sqrt{\frac{t+d}{n}}$ and $\max_{I\in J_d}|\sigma_{\min}-1| \lessapprox \sqrt{\frac{t+d}{n}}$.

Now let's replace t with $t+\log\mathrm{card}(J_d)$ and further change it to $t+d\log\frac{eN}{d}$.

Then with probability $\geq 1-e^{-t}$, $\max_{I\in J_d}|\sigma_{\max}-1| \lessapprox \sqrt{\frac{t+d+d\log\frac{eN}{d}}{n}} \lessapprox \sqrt{\frac{t+d\log\frac{eN}{d}}{n}}$, similarly, $\max_{I\in J_d}|\sigma_{\min}-1| \lessapprox \sqrt{\frac{t+d\log\frac{eN}{d}}{n}}$.

Now choose $t = d\log\frac{eN}{d}$, with probability $\geq 1-(\frac{eN}{d})^{-d}$, we have $\max_{I\in J_d}|\sigma_{\max}-1|\lessapprox\sqrt{\frac{d\log\frac{eN}{d}}{n}}$ and $\max_{I\in J_d}|\sigma_{\min}-1|\lessapprox\sqrt{\frac{d\log\frac{eN}{d}}{n}}$.

So we should take δ such that $\sqrt{1-\delta}\leq 1-c\sqrt{\frac{d\log\frac{eN}{d}}{n}}\leq\sigma_{\min}(X)\leq\sigma_{\max}(X)\leq 1+c\sqrt{\frac{d\log\frac{eN}{d}}{n}}\leq\sqrt{1+\delta}$.

In fact, it's enough to take $\delta\approx c'\sqrt{\frac{d\log\frac{eN}{d}}{n}}$, we then have with probability $\geq 1-(\frac{eN}{d})^{-d}$, $\delta_d(X)\lessapprox\sqrt{\frac{d\log\frac{eN}{d}}{n}}$. $\square$

Let's discuss **sparsity problems with noise**.

The model is $Y = X\beta_* + \xi, \beta_*\in\mathbb{R}^N$ where X is $n\times N$ design matrix, $\xi\sim N(0,\sigma^2 I_n)$ and $n << N$.

The error of LS is $\frac{N\sigma^2}{n}$.

Suppose β_* is sparse, or $d(\beta_*) = \sum_{i=1}^N \mathbb{I}(\beta_{i*}\neq 0) << N$.

One natural candidate to solve this problem is to use the penalized least square $||Y-X\beta||^2+\varepsilon d(\beta)$ and min this over $\beta\in\mathbb{R}^N$. A typical choice ε is σ^2. This is non-convex, not smooth, so not a good optimization.

This leads us to the convex relaxation.

Let $\hat\beta := \operatorname{argmin}_{\beta\in\mathbb{R}^N}\{||Y-X\beta||^2+\varepsilon||\beta||_{l_1}\}$ and a typical $\varepsilon = c\sqrt{\log N}$ and recall it's the **LASSO** estimator.

Proposition 5.7. For $J\subset\{1,\cdots,N\}$ and $b>0$, define $C_J^{(b)} = \{u\in\mathbb{R}^N : \sum_{j\notin J}|u_j|\leq b\sum_{j\in J}|u_j|\}$ and $\gamma^{(b)}(J,X) = \inf\{C>0 : \sum_{j\in J}|u_j|^2\leq C^2||Xu||^2, u\in C_J^{(b)}\}$. One can bound $\gamma^{(b)}(J,X)$ for J with $\text{card}(J)=d$ in terms of restricted isometry constants $\delta_{3d}(X)$, as in the case of $b=1$. For any $\beta\in\mathbb{R}^N$ with $J_\beta = \text{supp}(\beta)$, let $\gamma(\beta) := \gamma^{(5)}(J_\beta, X)$.

Definition 5.9. For $u\in\mathbb{R}^N$, we denote $||u||_{l_p} = (\sum_{i=1}^n|u_i^p|)^{\frac{1}{p}}$ for $p\geq 1$ and $||u||_{l_\infty} = \max_{1\leq i\leq n}|u_i|$.

Definition 5.10. Convex function

$f:\mathbb{R}^N\to\mathbb{R}$ is convex if and only if for all $x_1,x_2\in\mathbb{R}^N$ and all $\lambda\in[0,1]$, $f(\lambda x_1+(1-\lambda)x_2)\leq\lambda f(x_1)+(1-\lambda)f(x_2)$. A function is convex if and only if it's always above its support line.

Definition 5.11. Subgradient and subdifferential

A vectors $w\in\mathbb{R}^N$ is a subgradient of a function f at point x means that $f(y)-f(x)\geq\langle w, y-x\rangle$. In other words, $f(x)+\langle w,y-x\rangle$ is a support function. The subdifferential $f(x) = \{w\in\mathbb{R}^N : w \text{ is a subgradient of } f \text{ at } x\}$. One can show this set is a convex set. This can be viewed as a function $x\in\mathbb{R}^N\mapsto f(x)$. If f is differentiable, then $\partial f(x) = \{\nabla f(x)\}$. For example, suppose $f(x)=|x|, x\in\mathbb{R}$, then

$$\partial f(x) = \begin{cases} \{1\} & , x > 0 \\ [-1,1] & , x = 0 \\ \{-1\} & , x < 0 \end{cases}$$

Theorem 5.6. *Sum rule for subdifferentials (Moreau-Rockeffellar theorem)*

If $f_1, \cdots, f_k : \mathbb{R}^N \to \mathbb{R}$ are convex functions where we assume they are bounded. Then $(f_1(x) + \cdots + f_k(x)) = \partial f_1(x) + \cdots + \partial f_k(x)$. Where $+$ is the Minkowski sum defined in definition 2.1, $c_1 + \cdots + c_k = \{x_1 + \cdots + x_k, x_1 \in c_1, \cdots, x_k \in c_k\}$. For example, $f(x) = ||x||_{l_1} = \sum_{i=1}^n |x_i|$, then

$$\partial ||x||_{l_1} = \sum_{i=1}^n \partial |x_i| = \{u \in \mathbb{R}^N\}$$

where

$$\partial u_j = \begin{cases} \{1\} & , x_j > 0 \\ [-1,1] & , x_j = 0 \\ \{-1\} & , x_j < 0 \end{cases}$$

Proposition 5.8. Suppose $x \in \mathbb{R}^N$ is a minimal point of a convex function $f : \mathbb{R}^N \to \mathbb{R}$. Or $f(x) = \min_{y \in \mathbb{R}^N} f(y)$. Then $0 \in \partial f(x)$. The proof is trivial. Just note that $\forall y, f(y) - f(x) \geq 0 = \langle 0, y - x \rangle \implies 0 \in f(x)$.

Theorem 5.7. *Monontonicity of subdifferential*

For any points $x_1, x_2 \in \mathbb{R}^N$, for $\forall w_1 \in \partial f(x_1), w_2 \in \partial f(x_2)$. We have $\langle w_1 - w_2, x_1 - x_2 \rangle \geq 0$. When $N = 1$ and f is smooth, $(w_1 - w_2)(x_1 - x_2) \geq 0 \Leftrightarrow (f(x_1) - f(x_2))(x_1 - x_2) \geq 0$. We can define monontonicity in $\mathbb{R}^N$ in this way as well.

Theorem 5.8. *Suppose $\varepsilon \geq 3||X^T \xi||_\infty$, then*

$$||X\hat{\beta} - X\beta_*||^2 \leq \inf_{\beta \in \mathbb{R}^N} [||X\beta - X\beta_*||^2 + c\gamma(\beta)^2 d(\beta)\varepsilon^2]$$

*where $\hat{\beta} := \operatorname{argmin}_{\beta \in \mathbb{R}^N} \{||Y - X\beta||^2 + \varepsilon ||\beta||_{l_1}\}$, c is a numerical constant and $\gamma(\beta) := \gamma^{(5)}(J_\beta, X)$. This is called **sparsity oracle inequality**. Here nothing is random and ξ is fixed.*

Proof

Write $\mathscr{L}(\beta) = ||X\beta - Y||^2 + \varepsilon||\beta||_{l_1}$. And $\hat{\beta} = \operatorname{argmin}_{\beta\in\mathbb{R}^n}\mathscr{L}(\beta)$. Then $\mathscr{L}(\beta)$ is a convex function on $\mathbb{R}^N$.

Since $\hat{\beta}$ is a minimizer of $\mathscr{L}(\beta)$, we have $0 \in \partial\mathscr{L}(\hat{\beta})$.

Also, $\partial\mathscr{L}(\beta) = 2X^T(X\beta - Y) + \partial||\beta||_{l_1}$.

It follows that $0 \in \partial\mathscr{L}(\hat{\beta}) \implies \exists \hat{w} \in \partial||\hat{\beta}||_{l_1}$ such that $2X^T(X\hat{\beta} - Y) + \varepsilon\hat{w} = 0$.

First, multiply both sides by $\hat{\beta} - \beta$, so

$$\langle 2X^T(X\hat{\beta} - Y), \hat{\beta} - \beta\rangle + \varepsilon\langle\hat{w}, \hat{\beta} - \beta\rangle = 0$$

Suppose $w \in \partial||\beta||_{l_1}$. Specifically, let

$$w_j = \begin{cases} 1 & , \beta_j > 0 \\ 0 & , \beta_j = 0 \\ -1 & , \beta_j < 0 \end{cases}$$

we have

$$2\langle X\hat{\beta} - Y, X\hat{\beta} - X\beta\rangle + \varepsilon\underbrace{\langle\hat{w} - w, \hat{\beta} - \beta\rangle}_{\geq 0} = \varepsilon\langle w, \beta - \hat{\beta}\rangle$$

Let $Y = X\beta_* + \xi$, then $2\langle X\hat{\beta} - X\beta_*, X\hat{\beta} - X\beta\rangle + \varepsilon\langle\hat{w} - w, \hat{\beta} - \beta\rangle = \varepsilon\langle w, \beta - \hat{\beta}\rangle + 2\langle\xi, X\hat{\beta} - X\beta\rangle$.

Also, $2\langle X\hat{\beta} - X\beta^*, X\hat{\beta} - X\beta\rangle = ||X\hat{\beta} - X\beta_*||^2 + ||X\hat{\beta} - X\beta||^2 - ||X\beta - X\beta_*||^2$.

Now,

$$||X\hat{\beta} - X\beta_*||^2 + ||X\hat{\beta} - X\beta||^2 - ||X\beta - X\beta_*||^2 + \varepsilon\langle\hat{w} - w, \hat{\beta} - \beta\rangle = \varepsilon\langle w, \beta - \hat{\beta}\rangle + \underbrace{2\langle X^T\xi, \hat{\beta} - \beta\rangle}_{\leq 2||X^T\xi||_\infty||\hat{\beta} - \beta||_{l_1}}$$

which follows from $|\langle u, v\rangle| = |\sum_i u_i v_i| \leq \max_i |u_i| \sum_i |v_i| = ||u||_\infty ||v||_{l_1}$.

When $||X\hat{\beta} - X\beta_*||^2 + ||X\hat{\beta} - X\beta||^2 - ||X\beta - X\beta_*||^2 \leq 0$, we have $||X\hat{\beta} - X\beta_*||^2 \leq ||X\beta - X\beta_*||^2$, and we finish the proof.

When $||X\hat{\beta} - X\beta_*||^2 + ||X\hat{\beta} - X\beta||^2 - ||X\beta - X\beta_*||^2 > 0$, we need the following:

<u>**Claim:**</u> $\hat{\beta} - \beta \in C^{(5)}_{J_\beta}$.

Proof

First, drop the first term, which is non-negative, we wave

$$\varepsilon\langle\hat{w}-w,\beta-\hat{\beta}\rangle\leq\varepsilon\langle w,\beta-\hat{\beta}\rangle+2||X^T\xi||_\infty||\hat{\beta}-\beta||_{l_1}$$

$$\varepsilon\langle\hat{w}-w,\beta-\hat{\beta}\rangle=\varepsilon\sum_{j=1}^{N}\underbrace{(\hat{w}_j-w_j)(\hat{\beta}_j-\beta_j)}_{\geq 0}\geq\varepsilon\sum_{j\notin J_\beta}\hat{w}_j\hat{\beta}_j=\varepsilon\sum_{j\notin J_\beta}|\hat{\beta}_j|=\varepsilon\sum_{j\notin J_\beta}|\hat{\beta}_j-\beta_j|$$

since each element is a subdifferential.

Now, $\varepsilon\langle w,\beta-\hat{\beta}\rangle=\varepsilon\sum_{j\in J_\beta}w_j(\hat{\beta}_j-\beta_j)\leq\varepsilon\sum_{j\notin J_\beta}|\hat{\beta}_j-\beta_j|$.

Also,

$$2||X^T\xi||_\infty||\hat{\beta}-\beta\rangle||_{l_1}\leq\frac{2}{3}\varepsilon||\hat{\beta}-\beta||_{l_1}=\frac{2}{3}\varepsilon\sum_{j\in J_\beta}|\hat{\beta}_j-\beta_j|+\frac{2}{3}\varepsilon\sum_{j\notin J_\beta}|\hat{\beta}_j-\beta_j|$$

As a result,

$$\varepsilon\sum_{j\notin J_\beta}|\hat{\beta}_j-\beta_j|\leq\frac{5}{3}\varepsilon\sum_{j\in J_\beta}|\hat{\beta}_j-\beta_j|+\frac{2}{3}\varepsilon\sum_{j\notin J_\beta}|\hat{\beta}_j-\beta_j|$$

Therefore, $\frac{1}{3}\varepsilon\sum_{j\notin J_\beta}|\hat{\beta}_j-\beta_j|\leq\frac{5}{3}\varepsilon\sum_{j\in J_\beta}|\hat{\beta}_j-\beta_j|$. So $\hat{\beta}-\beta\in C^{(5)}_{J_\beta}$. □

Let's go back to the inequality and call $2\langle X^T\xi,\hat{\beta}-\beta\rangle$ as the main identity.

It follows from the main identity, the following is ture:

$$||X\hat{\beta}-X\beta_*||^2+||X\hat{\beta}-X\beta||^2+\varepsilon\sum_{j\notin J_\beta}|\hat{\beta}_j-\beta_j|\leq||X\beta-X\beta_*||^2+\varepsilon\sum_{j\in J_\beta}|\hat{\beta}_j-\beta_j|+\frac{2}{3}\varepsilon\sum_{j\in J_\beta}|\hat{\beta}_j-\beta_j|+\frac{2}{3}\varepsilon\sum_{j\notin J_\beta}|\hat{\beta}_j-\beta_j|$$

since $\varepsilon\sum_{j\notin J_\beta}|\hat{\beta}_j-\beta_j|$ is the lower bound of the main identity.

Therefore,

$||X\hat{\beta}-X\beta_*||^2+||X\hat{\beta}-X\beta||^2+\frac{1}{3}\varepsilon\sum_{j\in J_\beta}|\hat{\beta}_j-\beta_j|$

$\leq||X\hat{\beta}-X\beta_*||^2+\frac{5}{3}\varepsilon\sum_{j\in J_\beta}|\hat{\beta}_j-\beta_j|$

$\leq||X\hat{\beta}-X\beta_*||^2+\frac{5}{3}\varepsilon(\sum_{j\in J_\beta}|\hat{\beta}_j-\beta_j|^2)^{1/2}\sqrt{d(\beta)}$ by Cauchy-Schwarz.

Since $\hat{\beta}-\beta\in C^{(5)}_{J_\beta}$, it follows that $(\sum_{j\in J_\beta}|\hat{\beta}_j-\beta_j|^2)^{1/2}\leq\gamma^{(5)}(J_\beta,X)||X\hat{\beta}-X\beta||$ and we will use $\gamma(\beta)$ to represent $\gamma^{(5)}(J_\beta,X)$.

So

$$||X\hat{\beta} - X\beta_*||^2 + ||X\hat{\beta} - X\beta||^2 \leq ||X\beta - X\beta_*||^2 + \underbrace{\frac{5}{3\sqrt{2}}\varepsilon\gamma(\beta)\sqrt{d(\beta)}}_{a}\underbrace{||X\hat{\beta} - X\beta||\sqrt{2}}_{b}$$

Using $ab \leq \frac{a^2+b^2}{2}$, we have

$$||X\hat{\beta} - X\beta_*||^2 + ||X\hat{\beta} - X\beta||^2 \leq ||X\beta - X\beta_*||^2 + \frac{5^2}{3^2 \cdot 2 \cdot 2}\varepsilon^2\gamma^2(\beta)d(\beta) + ||X\hat{\beta} - X\beta||^2$$

So

$$||X\hat{\beta} - X\beta_*||^2 \leq ||X\beta - X\beta_*||^2 + \underbrace{\frac{5^2}{3^2 \cdot 2^2}}_{c}\varepsilon^2\gamma^2(\beta)d(\beta)$$

□

Corollary 5.9. *Take $\beta = \beta_*$, we get $||X\hat{\beta} - X\beta_*||^2 \leq c\gamma(\beta_*)^2 d(\beta_*)\varepsilon^2$.*

Note that X is $n \times N$, X^T is $N \times n$, so $X^T\xi \in \mathbb{R}^N$. Pick canonical bases of $\mathbb{R}^N : e_1, \cdots, e_N$, then $||X^T\xi||_\infty = \max_{1\leq j\leq N}|\langle X^T\xi, e_j\rangle| = \max_{1\leq j\leq N}|\langle \xi, X^T e_j\rangle|$. Let $x_j = X^T e_j$ be the $j-$th column of X, then $||X^T\xi||_\infty = \max_{1\leq j\leq N}|\langle \xi, e_j\rangle|$.

Note that $\langle \xi, e_j\rangle \sim N(0, \sigma^2||x_j||^2)$, so $\mathbb{P}(\langle \xi, e_j\rangle \geq \sigma||x_j||\sqrt{t}) \leq 2e^{-t/2}$.

Therefore, $\mathbb{P}(||X^T\xi||_\infty \geq \sigma\max_{1\leq j\leq N}||x_j||\sqrt{t}) \leq 2Ne^{-t/2}$ by the union bound.

Let $t \to t + 2\log N$, then $\mathbb{P}(||X^T\xi||_\infty \geq \sigma\max_{1\leq j\leq N}||x_j||\sqrt{t + 2\log N}) \leq 2e^{-t/2}$.

Let's now assume that $\varepsilon \geq 3\sigma\max_{1\leq j\leq N}||x_j||\sqrt{t + \log N}$, then with probability $\geq 1 - 2e^{-t/2}$, we have $\varepsilon \geq 3||X^T\xi||_\infty$.

Theorem 5.10. *Assume that $\varepsilon \geq 3\sigma\max_{1\leq j\leq N}||x_j||\sqrt{t + 2\log N}$, then with probability at least $1 - 2e^{-t/2}$, the following bound holds: $||X\hat{\beta} - X\beta_*||^2 \leq \inf_{\beta\in\mathbb{R}^N}[||X\beta - X\beta_*||^2 + c\gamma(\beta)^2 d(\beta)\varepsilon^2]$. In particular,*

$$||X\hat{\beta} - X\beta_*||^2 \leq c\varepsilon^2\gamma^2(\beta_*)d(\beta_*) \leq \max_{1\leq j\leq N}||x_j||^2(t + \log 2N)\gamma^2(\beta_*)\sigma^2 d(\beta_*)$$

Let's now talk about some trace regression models examples.

1. Matrix completion (Netflix) problem

Let A be $m \times m$ matrix, (could be $m_1 \times m_2$, but for simplicity let's assume it's square for now).

The complexity is how do we consider this problem. For vector, we use sparsity. We will use rank for matrix.

Suppose A is symmetric, then we have $A = \sum_{j=1}^{r} \lambda_j(\phi_j \otimes \phi_j)$ where $\lambda_j \neq 0$ and r is the rank of the matrix A. For r eigenvectors, we need $r \times m$ for these eigenvectors and r for eigenvalues. So need about rm numbers to represent this matrix A. If $r << m$, we can let r be the number of freedom in this matrix problem. Note that we need about m^2 for a general symmetric matrix. A natural question is that whether we can recover a matrix with low-rank r and observations $< r$. Consider the matrix with one element 1 and 0 elsewhere. Then the probability we are missing this element is $(1-\frac{1}{m})^n$ and we need $n = o(m^2)$ to recover the matrix.

2. Quantum State Tomography

Density matrix $\rho : \mathbb{C}^m \to \mathbb{C}^m$ is a $m \times m$ Hermitian (self-adjoint) matrix in the Hilbert space. Assume ρ is positive semi-definite. The assumption is $tr(\rho) = 1$, like $\int_{\mathbb{R}} f(x)dx = 1$.

Observables are represented by Hermitian (self-adjoint) $m \times m$ matrix.

Suppose X is an observable, we want to measure X in state ρ. Then $X = \sum_j \lambda_j P_j, P_j = \phi_j \otimes \phi_j$ where ϕ_j are eigenvectors and $\lambda_j \in \mathbb{R}$.

Let Y be the value of the observable X in state ρ, then $\mathbb{P}_\rho(Y = \lambda_j) = tr(\rho P_j) \geq 0, j = 1, \cdots, m$. Then $\mathbb{E}_\rho Y = \sum_j \lambda_j tr(\rho P_j) = tr(\rho \sum_j \lambda_j P_j) = tr(\rho X)$.

Let $X_1, \cdots, X_n$ be observables and (by physicists), and n copies of quantum system are prepared in state ρ (this is often difficult to do). Let $Y_1, \cdots, Y_n$ be the values of $X_1, \cdots, X_n$. The goal of **quantum state tomography** is to estimate ρ based on $(X_1, Y_1), \cdots, (X_n, Y_n)$.

Recall that $\mathbb{E}_\rho(Y_j|X_j) = tr(\rho X_j)$, then we can write $Y_j = tr(\rho X_j) + \xi_j$ where $\mathbb{E}[\xi_j|X_j] = 0$. This is similar to linear regression. And the matrix is usually high-dimension, but they can often be **approximated** by low-rank matrix, since physicists often try to prepare system in pure states.

Definition 5.12. Trace regression model

The model is $Y_j = tr(\rho X_j) + \xi_j$ where ρ is the target matrix, Y is response and ξ_j is noise. We are assuming that ρ is low-rank, or can be well approximated by low-rank matrices.

Definition 5.13. Nuclear norm

$||\rho||_1 = tr(\sqrt{\rho^2}) = \sum_{j=1}^{m} |\lambda_j(\rho_j)|$, it's the sum of singular values for rectangle matrices.

A typical method is called the **matrix LASSO**. Let

$$\hat{\rho} = \text{argmin}_{\rho \in \mathbb{H}_m}[\frac{1}{n}(Y_i - \langle \rho, X_i \rangle)^2 + \varepsilon ||\rho||_1]$$

and we can show that

$$\frac{1}{m^2}||\hat{\rho} - \rho||_2^2 \lessapprox \frac{\sigma_\xi^2 m \, rank(\rho)}{n} \log(factor)$$

where we are using the Hilbert-Schmidt norm, and it's similar to what we had before.