STAT 230 - MULTIVARIATE STATISTICAL ANALYSIS

GREG TAM

Contents

1. September 3rd, 2013

In the 1950's, we typically had n large and p small, where n is the number of subjects/individuals/samples and p is the number of dimensions of features such as blood pressure/weight/height. In big data/massive data/high dimensional data, we have n which is large or small, but p is very large (e.g. $n = 50$, $p = 7000$.)

A: Matrix Theory.

A1: Vector Space. A1.1:
$\mathbb{R}^p$: Euclidean Space
$x \in \mathbb{R}^p$ means that

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

Inner Product:

$$\langle x, y \rangle = \sum_{i=1}^{p} x_i y_i$$

a) Cauchy-Schwarz

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$$

b) If $\alpha_{x,y}$ is the angle between x and y, then

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

$$0 \leq \alpha_{x,y} \leq \pi \Rightarrow \alpha_{x,y} = \cos^{-1}\left(\frac{\langle x, y \rangle}{\|x\| \cdot \|y\|}\right)$$

c) Given A is a $p \times p$ positive definite matrix (e.g. $A = I_p$), the general form of the inner product is

$$\langle x, y \rangle = \underbrace{x'}_{1\times p}\underbrace{A}_{p\times p}\underbrace{y}_{p\times 1}$$

Cauchy-Schwarz continues to hold.

A1.2 Linear Space. Let

$$\underbrace{V}_{p\times q} = \begin{bmatrix} v_1 & v_2 & \cdots & v_q \end{bmatrix}$$

where $1 \leq q \leq p$. Then we let

$$L_{col}(V) = \left\{ \sum_{i=1}^{q} a_i v_i, \text{for all } a = \begin{pmatrix} a_1 \\ \vdots \\ a_q \end{pmatrix} \right\}$$

For simplicity, $v_1, \ldots, v_q$ are linearly independent. Otherwise, we can find a subset which spans the same space.

(a) If $v_1, \ldots, v_q$ are linearly independent, they form a basis of $L_{col}(V)$

(b) Sometimes you want an orthogonal basis

$$\underbrace{V}_{p\times q} = \underbrace{\Gamma}_{p\times q}\underbrace{C}_{q\times q}$$

which is the Q-R decomposition. We have that

$$\Gamma = \begin{bmatrix} \gamma_1 & \gamma_2 & \ldots & \gamma_q \end{bmatrix}$$

and $\Gamma'\Gamma = I_q$, $\Gamma\Gamma' \neq I_p$.

- Γ has orthogonal columns
- C is upper triangular
- If $v_1, \ldots, v_q$ are linearly independent, then the diagonals of C are non-zero

We claim $L_{col}(V) = L_{col}(\Gamma)$. If $v \in L_{col}(V)$, then

$$\begin{aligned} v &= Vx \qquad \text{for some } x \\ &= \Gamma \frac{cx}{y} \\ &= \Gamma y \end{aligned}$$

which implies $v \in L_{col}(\Gamma)$. Similarly, since $\Gamma = VC^{-1}$, using the same argument, then for any $v \in L_{col}(\Gamma)$, we have $v \in L_{col}(V)$

(c) Orthocomplement

$$\begin{aligned} L_{col}^{\perp} &= \text{set of all vector orthogonal to v} \\ &= \{y \in \mathbb{R}^p : \langle y, v \rangle = 0, \text{ for all } v \in L_{col}(V)\} \end{aligned}$$

So we have

$$V = L_{col}(V) + L_{col}^{\perp}(V)$$

which means

$$W \subseteq V \Rightarrow W^{\perp} \supseteq V^{\perp}$$

Homework (A1.1)**.** *Show that $y \in L_{col}^{\perp}(V)$ if and only if $\langle y_i, v_i \rangle = 0$, $i = 1, 2, \ldots, q$*

Theorem 1 (Cauchy)**.** *Given $x = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}$ then $L_{col}(\underbrace{X}_{p\times n}) = L_{col}(XX')$*

Proof. "$\subseteq$" + "$\supseteq$". The direction $\supseteq$ is trivial. To show $\subseteq$, it is sufficient to show

$$L_{col}^{\perp}(X) \supseteq L_{col}^{\perp}(XX')$$

To this end, let us say $v \in L_{col}^{\perp}$, then $XX'v = 0$ and so

$$v'(XX')v = 0 \Leftrightarrow \|X'v\|^2 = 0 \Rightarrow X'v = 0$$

which implies $v \in L_{col}^{\perp}(X)$. □

Rank: $X = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}, x_i \in \mathbb{R}^p$

- Rank(X) = maximum # of linearly independent columns of X
- Rank(X) = maximum number of linearly independent rows of X
- Rank(X) =Rank(X')
- If B and C are nonsingular, then

$$\text{Rank}(BXC) = \text{Rank}(X)$$

- $\text{Rank}(X) = \dim(L_{col}(X))$
- By the above and Cauchy's theorem,

$$\dim(L_{col}(X)) = \text{Rank}(X) = \text{Rank}(XX') = \dim(L_{col}(XX'))$$

Gram matrix: X, which is $p \times n$. Row: XX' which is $p \times p$.

- symmetric
- positive definite
-

$$\begin{aligned} G &= \underbrace{\Gamma}_{p\times r}\underbrace{D}_{r\times r}\underbrace{\Gamma'}_{r\times p} \\ &= \sum d_i \gamma_i \gamma_i' \end{aligned}$$

where D is the diagonal matrix with entries $d_1, \dots, d_r$, $r = \text{Rank}(X)$, $\Gamma = \begin{bmatrix}\gamma_1 \dots \gamma_r\end{bmatrix}$, $\Gamma'\Gamma = I_r$.

A1.3 $\mathcal{L}^2$-space. Suppose

$$Z(w) = \begin{pmatrix} z_1(w) \\ z_2(w) \\ \vdots \\ z_p(w) \end{pmatrix}$$

is a p-dimensional random variables on $\mathcal{L}^2(\Omega, \mathfrak{B}, \mathcal{P})$

- For simplicity, let $\mathbb{E}[Z] = 0$, let the covariance

$$\begin{aligned} \text{Cov}(Z) &= \mathbb{E}[ZZ'] \\ &= \Sigma \end{aligned}$$

Linear space

$$\begin{aligned} L(Z) &= \{X = x'Z, x \in \mathbb{R}^p\} \\ &= \sum_{i=1}^{p} x_i z_i \\ &= \text{sub-space of linear functions of } \mathbb{Z} \end{aligned}$$

- $\mathbb{E}[X] = \mathbb{E}[X'Z] = X'\mathbb{E}[Z] = 0$
- Inner product

$$\begin{aligned} \langle X, Y\rangle &= \text{Cov}(X, Y) \\ &= \text{Cov}(x'Z, y'Z) \\ &= x'\Sigma y \end{aligned}$$

- Length: $\|X\|^2 = \langle X, X\rangle$
- Angle: $\cos(x, y) = \frac{\langle X,Y\rangle}{\|X\|\cdot\|Y\|}$

Homework (A1.3)**.** $\Omega = \{1, 2, \cdots, L\}$, $p = (\pi_1, \pi_2, \dots \pi_L)$, $\pi_i \geq 0$, $\sum \pi_i = 1$

$$Z = \begin{pmatrix} z_1 \\ \vdots \\ z_L \end{pmatrix} \qquad z_i = \mathbb{1}\{w = i\} = \begin{cases} 1 & \pi_i \\ 0 & 1 - \pi_i \end{cases}$$

Give an explicit formula for $\langle X, Y\rangle$.

Homework (A1.4)**.** *$\mathcal{L}(Z)$ is a linear space of $\mathcal{L}^2(\Omega, \mathfrak{B}, \mathcal{P})$. Argue that $\mathcal{L}^{\perp}(Z)$ is the set of all random variables uncorrelated with Z.*

2. September 9th, 2013

2.1. **Projections, Gram-Schmidt Orthogonalization, Triangularization.** Let $V = (v_1, \ldots, v_p)$ with $v_i \in \mathbb{R}^n$, $1 \leq p \leq n$, $\text{rank}(V) \equiv \dim(L_{col}(V)) = p$

We wish that for any $y \in \mathbb{R}^n$, $y = \hat{y} + \overset{\perp}{y}$ such that $\hat{y} \in L_{col}(V)$ and $\overset{\perp}{y} \in L_{col}^{\perp}(V)$ $\hat{y} \in L_{col}(V)$ Therefore, we have $\hat{y} = V\hat{\beta}$, for some $\hat{\beta} \in \mathbb{R}^p$. Now,

$$\begin{aligned} V'y &= V'(\hat{y} + \overset{\perp}{y}) \\ &= V'[V\hat{\beta} + \overset{\perp}{y}] \\ &= \underbrace{V'V}_{p\times p}\hat{\beta} \end{aligned}$$

which implies

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

So then

$$\hat{y} = V\hat{\beta} = V(V'V)^{-1}V'y \equiv \hat{p}y$$

and

$$\begin{aligned} \overset{\perp}{y} &= y - \hat{y} \\ &= [I_n - V(V'V)^{-1}V']y \\ &= \overset{\perp}{p}y \qquad (\overset{\perp}{p} \equiv I_n\hat{p}) \\ &\equiv [I_n - \hat{p}]y \end{aligned}$$

Where

$$\begin{aligned} \hat{p} &: \mathbb{R}^n \to L_{col}(V) \\ \overset{\perp}{p} &: \mathbb{R}^n \to L_{col}^{\perp}(V) \end{aligned}$$

Facts:

(a) for any $x \in L_{col}(V)$, $\hat{p}x = x$, $x = \hat{x} + \overset{\perp}{x} \in L_{col}^{\perp}(V)$, $x \in L_{col}^{\perp}(V)$, $\hat{p}x = 0$. If $x \in L_{col}(V)$, then $x = Vy$, for some $y \in \mathbb{R}^p$.

$$\begin{aligned} \hat{p}x &= \hat{p}Vy \\ &= V(V'V)^{-1}V'(Vy) \\ &= Vy \\ &\equiv x \end{aligned}$$

(b) Idempotent means that $\hat{p}^2 = \hat{p}$, which means that for all x,

$$\begin{aligned} \hat{p}^2x &= \hat{p}(\hat{p}x) \\ &= \hat{p}x \end{aligned}$$

(c)

$$L_{col}(\hat{p}) = L_{col}(V)$$

"$\subseteq$" is trivial. To prove the other direction, we have that for any $x \in L_{col}(V)$, $\hat{p}x = x \Rightarrow x \in L_{col}(\hat{p})$

(d)

$$\begin{aligned} \text{rank}(\hat{p}) &= p \\ \text{rank}(\overset{\perp}{p}) &= n - p \end{aligned}$$

(e)

$$\begin{aligned} \|y\|^2 &= \|\hat{y}\|^2 + \|\overset{\perp}{y}\|^2 \\ &\equiv y'\hat{p}y + y'(I_n - \hat{p})y \end{aligned}$$

(f) If $V = \Gamma M$ where V and Γ are $n \times p$ and M is $p \times p$, $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_p)$, $\Gamma'\Gamma = I_p$, then

$$\begin{aligned} \hat{p} &= V(V'V)^{-1}V' \\ &= (\Gamma M)[M'\Gamma'\Gamma M]^{-1}M'\Gamma' \\ &= \Gamma(M(M'M)^{-1}M')\Gamma' \\ &= \Gamma\Gamma' \\ &\equiv \Gamma\cancel{(\Gamma'\Gamma)^{-1}}\Gamma' \end{aligned}$$

(g) eigenvalues of $\hat{p}$ are either 0 or 1

(h)

$$\|\hat{\tilde{y}}\|^2 = \min_{v \in L_{col}(V)} \|y - v\|^2$$

(i)

$$\cos^2(\alpha_{y,\hat{y}}) = \frac{\|\hat{y}\|^2}{\|y\|^2}$$

and

$$\alpha_{y,\hat{y}} = \min_{v \in L_{col}(v)} \alpha_{y,v}$$

If λ is an eigenvalue, then $\hat{p}x = \lambda x$ for some x. If $\lambda \neq 0$, then

$$x = \frac{1}{\lambda}\hat{p}x \in L_{col}(V) \Rightarrow \hat{p}x = x \Rightarrow \lambda = 1$$

Homework (A1.2)**.** *Suppose $L_{col}(V) = L_{col}(\Gamma)$ and $R = \Gamma\Delta\Gamma'$ where Δ is $p \times p$ and orthogonal ($\Delta'\Delta = I_p$). Show R is a rotation matrix in $L_{col}(V)$ i.e. for any $v \in L_{col}(V)$,*

$$\tilde{v} \equiv Rv \in L_{col}(V)$$

and

$$\|\tilde{v}\| = \|v\|$$

Example 1. *$v = \{\mathbf{1}\}$, where $n = n$ and $p = 1$. This is an $n \times 1$ vector of ones.*

$$\begin{aligned}\hat{p} &= V(V'V)^{-1}V' \\ &= \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' \\ &= \frac{1}{n}\mathbf{1}\mathbf{1}'\end{aligned}$$

$$\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}\begin{pmatrix}1 & \cdots & 1\end{pmatrix} = \begin{pmatrix}1 & \cdots & 1\\ \vdots & & \vdots \\ 1 & \cdots & 1\end{pmatrix}$$

Now, for any $y \in \mathbb{R}^n$,

$$\begin{aligned}\hat{p}y &= \frac{1}{n}\mathbf{1}\mathbf{1}'y = \bar{y}\mathbf{1} \\ \hat{\tilde{p}}y &= y - \bar{y} \cdot 1 \\ \|\hat{p}y\|^2 &= \sum (y_i - \bar{y})^2\end{aligned}$$

Two familiar uses are

- Noise Removal: $\bar{y}$ for estimating means, e.g. $y_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$
- Signal Removal: so that we can study the noise, say variance of noise

A2.2 Projection in general form. Let A, which is $n \times p$ positive definite ($A > 0$)

Definition 1. *Let $\langle x, y\rangle = x'Ay$*

Now, if you wish to generalize projections to such a setting, $y = \hat{y} + \hat{\tilde{y}}$

$$\langle \hat{y}, \hat{\tilde{y}}\rangle = 0 \equiv \hat{y}A\hat{\tilde{y}} = 0$$

$$\begin{aligned}\hat{y} &= \underbrace{(V(V'AV)^{-1}V')}_{\hat{p}} \cdot A \cdot y \\ &\equiv \hat{p}Ay \\ \hat{\tilde{y}} &= \underbrace{(A^{-1} - V(V'AV)^{-1}V')}_{\hat{\tilde{p}}} \cdot A \cdot y \\ &\equiv \hat{\tilde{p}}Ay\end{aligned}$$

where $\hat{p} + \hat{\tilde{p}} = A^{-1}$.

Facts:

(a) $\hat{p}A\hat{p} = \hat{p}$

(b)

$$\begin{aligned}\|\hat{y}\|^2 &= \|\hat{p}Ay\|^2 \\ &= y'A\hat{p}A\hat{p}Ay \\ &= y'A\hat{p}Ay \\ &= y'AV(V'AV)^{-1}V'Ay \\ &= \langle y, V\rangle(\langle V, V\rangle)^{-1}\langle V, y\rangle\end{aligned}$$

Note: If A is $n \times k$ and B is $n \times l$, then $\langle A, B\rangle$ denotes a $k \times l$ matrix.

(c) $\|\overset{\perp}{y}\|^2 = \langle y, y\rangle - \langle y, V\rangle(\langle V, V\rangle)^{-1}\langle V, y\rangle$

Homework (A2.2)**.** *verify (a) - (c)*

A2.3 Application to linear prediction. Given $Y, v_1, \ldots, v_p$ with

- All are univariate random variables on the same $\mathcal{L}^2(\Omega, \beta, p)$
- Means are all 0
- Variances are finite

Goal: Find a linear combination of v_i to correlate with Y as much as possible.

Think of Ω as finite, Y as a vector

$$\boldsymbol{Y} \equiv \begin{pmatrix} \vdots \\ Y(\omega) \\ \vdots \end{pmatrix}$$

$\boldsymbol{V} = (v_1, \ldots, v_p)$ as a matrix

$$\boldsymbol{V} = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ v_1(\omega) & v_2(\omega) & \ldots & v_p(\omega) \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

where $\omega \in \Omega$.

Define the inner product as

$$\begin{aligned}\langle y, v\rangle &= \mathrm{Cov}(y, v) \\ &= (\mathrm{Cov}(y, v_1) \ldots \mathrm{Cov}(y, v_p)) \\ &= \mathbb{E}[y'\boldsymbol{V}]\end{aligned}$$

Now,

$$\begin{aligned}\hat{\boldsymbol{y}} &= \boldsymbol{V}\langle \boldsymbol{V}, \boldsymbol{V}\rangle^{-1}\langle \boldsymbol{V}, y\rangle \\ &= \boldsymbol{V}(\sigma_{V,V}^{-1} \cdot \sigma_{V,y})\end{aligned}$$

where

$$\begin{aligned}\Sigma &= \mathrm{Cov}\left((v_1, \ldots, v_p, y)\right) \\ &= \begin{bmatrix} \sigma_{V,V} & \sigma_{V,y} \\ \sigma_{y,V} & \sigma_{y,y} \end{bmatrix}\end{aligned}$$

$$\hat{\boldsymbol{y}} = \begin{pmatrix} \vdots \\ \hat{y}(\omega) \\ \vdots \end{pmatrix} = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ V_1(\omega) & V_2(\omega) & \ldots & V_p(\omega) \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \sigma_{V,V}^{-1}\sigma_{V,y}$$

which is the same as

$$\hat{y} = (\sigma_{y,V}\sigma_{VV}^{-1}) \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_p \end{pmatrix}$$

which is a linear combination of v_i. $y = \hat{y} + \overset{\perp}{y}$.

Facts:

(a) $\hat{y}$ is uncorrelated with $\overset{\perp}{y}$ or equivalently $\langle \hat{y}, \overset{\perp}{y}\rangle = 0$

(b) $\mathrm{Var}(\hat{y}) = \sigma_{yV}\sigma_{VV}^{-1}\sigma_{Vy}$

$$\mathrm{Var}(\hat{y}) = \|\overset{\perp}{y}\|^2 = \langle y, V\rangle\langle V, V\rangle^{-1}\langle V, y\rangle$$

(c) $\mathrm{Var}(\overset{\perp}{y}) = \sigma_{YY} - \sigma_{YV}\sigma_{VV}^{-1}\sigma_{VY}$

(d)

$$\frac{\mathrm{Var}(\hat{y})}{\mathrm{Var}(y)} = \frac{\sigma_{YV}\sigma_{VV}^{-1}\sigma_{VY}}{\sigma_{YY}}$$

Homework (A2.4)**.** *$\hat{y}$ is the linear combination of $v_1, \ldots, v_p$ that is most correlated with y, i.e.*

$$\mathrm{Corr}(y, \hat{y}) \geq \mathrm{Corr}(y, w)$$

where w is a linear combination of v_i's.

A2.4 Gram-Schmidt Orthogonalization. $V = (v_1, v_2, \ldots, v_p)$ with rank p. Define $V_j = (v_1, v_2, \ldots, v_j)$ which is $n \times j$, where for $j = 1, 2, \ldots, p$, $\hat{p}_j = V_j(V_j'V_j)^{-1}V_j'$.

	$w_1 = \frac{v_1}{\|v_1\|}$
$\overset{\perp}{v}_2 = (I - \hat{p}_1)v_2$	$w_2 = \frac{\overset{\perp}{v}_2}{\|\overset{\perp}{v}_2\|}$
$\overset{\perp}{v}_3 = (I - \hat{p}_2)v_3$	$w_3 = \frac{\overset{\perp}{v}_3}{\|\overset{\perp}{v}_3\|}$
$\overset{\perp}{v}_p = (I - \hat{p}_{p-1})v_p$	$w_p = \frac{\overset{\perp}{v}_p}{\|\overset{\perp}{v}_p\|}$

- $L_{col}(w_1, \ldots, w_j) = L_{col}(v_1, \ldots, v_j)$
- $w_j \perp L_{col}(v_1, \ldots, v_j) \equiv L_{col}(v_1, \ldots, v_{j-1})$
- $W'W = WW' = I_p$

Moreover,

$$\begin{aligned} v_1 &= t_{11}w_1 \\ v_2 &= t_{12}w_1 + t_{22}w_2 \\ &\vdots \\ v_p &= t_{1p}w_1 + t_{2p}w_2 + \cdots + t_{pp}w_p \end{aligned}$$

Now, $t_{ij} = (w_i, v_i)$. Especially,

$$\begin{aligned} t_{ii} &= (w_i, v_j) \\ &= \left(\frac{\overset{\perp}{v}_i}{\|\overset{\perp}{v}_i\|}\right)' v_i \\ &= \frac{\overset{\perp}{v}_i' v_i}{\|\overset{\perp}{v}_i\|} \qquad\qquad = \frac{\|\overset{\perp}{v}_i\|^2}{\|\overset{\perp}{v}_i\|} > 0 \end{aligned}$$

$$V = W \begin{bmatrix} t_{11} & t_{12} & \ldots & t_{1p} \\ 0 & t_{22} & \ldots & t_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & t_{pp} \end{bmatrix} \equiv WT$$

Homework (A2.5)**.**

(a) *Show that w_j is a linear combination of $v_1, \ldots, v_p$, so $W = VT^{-1}$*

(b) *Show in general, inverses of an upper triangular matrix is also upper triangular*

3. September 11th, 2013

Last time:

- Projections/basic facts
- Application to Gram-Schmidt
- Projection in general forms
- Application to linear prediction

For today:

- Determinants/volumes
- Spectral decomposition
- SVD (Singular Value Decomposition)
- Pseudo-inverse

A3.1 Determinant.

Definition 2 (Classical Definition)**.** *if A is a $p \times p$ matrix with $A = (a_{ij})_{1 \le i,j \le p}$, then*

$$|A| = \sum_{\pi} \operatorname{sgn}(\pi) a_{1\pi_1} a_{2\pi_2} \dots a_{p\pi_p}$$

where $\pi = (\pi_1, \dots, \pi_p)$ is a permutation of $(1, 2 \dots, p)$.

Facts:

(a) If T is triangular (upper or lower),

$$|T| = \prod_{i=1}^{p} t_{jj}$$

(b) $|A| = |A'|$

(c)

$$|AB| = |A||B| \Rightarrow |A^{-1}| = \frac{1}{|A|}$$

(d) If $\operatorname{rank}(A) < p$, $|A| = 0$

(e)

$$|(a_1, \dots, a_j + a_{j'}, a_{j+1}, \dots, a_p)| = |A|$$

with $j' \neq j$. This says if you add any column to another column, the determinant doesn't change

$$|(a_1, \dots, ca_j, a_{j+1}, \dots, a_p)| = c|A|$$

(f)

$$\left| \begin{bmatrix} A & \mathbf{0} \\ C & B \end{bmatrix} \right| = |A||B|$$

This is a block matrix with A being $p_1 \times p_1$ and B being $p - p_1 \times p - p_1$.

Homework (A3.1)**.** *If Γ is $p \times p$ orthogonal, then $|\Gamma| = \pm 1$.*

3.1. **Volume Interpretation.** $A = (a_1, \dots, a_p)$ full rank. Write $A = WT$ where T is upper triangular. Recall that

$$t_{jj} = \|a_j^{\perp}\|$$

the length of a_j after projected to $L(a_1, \dots, a_{j-1})$

$$|A| = |W||T| = \pm \prod_{j=1}^{p} \|a_j^{\perp}\|$$

Let us consider the easy case $p = 2$, then

$$a_1^{\perp} = a_1$$
$$a_2^{\perp} = a_2 - \frac{(a_1, a_2)}{\|a_1\|^2} a_1$$

$$VOLUMEDIAGRAM$$
$$|T| = \|a_1\| \cdot \|a_2^{\perp}\|$$

Claim.

$$|A| = \pm \cdot \textit{Volume of the parallepiped formed by vector } a_1, \dots, a_p$$

(g)

$$\left| \begin{bmatrix} A & B \\ C & D \end{bmatrix} \right| = \begin{cases} |D| \cdot |A - BD^{-1}C| & \text{if } \operatorname{rank}(D) = p - p_1 \\ |A| \cdot |D - CA^{-1}B| & \text{if } \operatorname{rank}(A) = p_1 \end{cases}$$

where A is $p_1 \times p_1$ and D is $p - p_1 \times p - p_1$.

$$\begin{aligned} \left| \begin{pmatrix} A & B \\ C & D \end{pmatrix} \right| &= \left| \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} I_{p_1} & \mathbf{0} \\ -D^{-1}C & I_{p-p_1} \end{pmatrix} \right| \\ &= \begin{pmatrix} A - BD^{-1}C & B \\ \mathbf{0} & D \end{pmatrix} \\ &= |D| \cdot |(A - BD^{-1}C)| \end{aligned}$$

(h)

$$|I_p + AB| = |I_n + BA|$$

where B is $n \times p$ and A is $p \times n$.

$$\begin{pmatrix} I_p & A \\ B & I_n \end{pmatrix} = \begin{cases} |I_p + AB| \\ |I_n + BA| \end{cases}$$

Homework (A3.2)**.** *Find*

$$\begin{vmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & 1 & \rho \\ \rho & \dots & \rho & 1 \end{vmatrix}$$

a matrix with 1 *along the diagonal and* ρ *everywhere else.*

A3.2 Spectral Decomposition. A : real/symmetric, $\text{rank}(A) = \gamma$, $A = p \times p$.

$$A = \underbrace{\Gamma}_{p\times r}\underbrace{\Lambda}_{r\times r}\underbrace{\Gamma'}_{r\times p} = \sum_{i=1}^{\gamma} \lambda_i \gamma_i \gamma_i'$$

where $\Gamma = (\gamma_1, \dots, \gamma_r)$ and $\Gamma'\Gamma = I_r$ and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_r)$. We have that

- λ_i's are the eigenvalues
- γ_i's are the eigenvectors

The proof is simple since we have

$$A = \Gamma\Lambda\Gamma'$$

so

$$A\Gamma = \Gamma\Lambda$$
$$A\Gamma_i = \lambda_i \gamma_i$$

Homework (A3.3)**.** *Show* $L_{col}(A) = L_{col}(\Gamma)$

Definition 3 (Trace)**.**

$$\text{tr}(A) = \sum_{i=1}^{p} a_{ii}$$

Fundamental Property:

$$\text{tr}(XY) = \text{tr}(YX)$$

where X *is* $p \times n$ *and* Y *is* $n \times p$

Example 2.

$$\begin{aligned} \text{tr}(\text{Cov}(X, X)) &= \text{tr}(\mathbb{E}[XX']) \\ &= \mathbb{E}[\text{tr}(XX')] \\ &= \mathbb{E}[\text{tr}(X'X)] \end{aligned}$$

Proof of Fundamental property. Write $Z = Y'$ so that X and Z have the same sizes

$$\begin{aligned} \text{tr}(XY) &= \text{tr}(XZ') \\ &= \sum_{i=1}^{p} \left(\sum_{j=1}^{n} X_{ij} Z_{ij} \right) \\ &= \sum_{i,j} X_{ij} Z_{ij} \\ &= (\boldsymbol{X}, \boldsymbol{Z}) \\ &= (\boldsymbol{Z}, \boldsymbol{X}) \\ &= \text{tr}(Z'X) \end{aligned}$$

$\boldsymbol{X}$ is the concatenation of all columns of X to make a vector of $\mathbb{R}^{pn}$. □

(a) A is symmetric, $A = \Gamma\Lambda\Gamma'$.

$$\begin{aligned}\operatorname{tr}(A) &= \operatorname{tr}(\Gamma\Lambda\Gamma') \\ &= \operatorname{tr}(\Lambda\Gamma\Gamma') \\ &= \operatorname{tr}(\Lambda) \\ &= \sum_{i=1}^{r} \lambda_i\end{aligned}$$

(b)

$$A = \Gamma(\Lambda^{1/2})\Gamma'\Gamma(\Lambda^{1/2})\Gamma'$$

A3.3 Singular Value Decomposition (SVD). X $n \times p$, $\operatorname{rank}(X) = r$, $p \leq n$ SVD:

$$\underbrace{X}_{n\times p} = \underbrace{U}_{n\times r}\underbrace{D}_{r\times r}\underbrace{V'}_{r\times p}$$

$$U'U = V'V = I_r$$

$$UU' \neq \text{Identity}$$

$$VV' \neq \text{Identity (except } r = p)$$

Let $D = \operatorname{diag}(d_1, \ldots, d_r)$ with $d_1 \geq d_2 \geq \ldots \geq d_r > 0$.

- d_i are called singular values
- columns of U: left singular vectors
- columns of V: right singular vectors

Say

$$\begin{aligned}X'X &= (VDU')(UDV') \\ &= VD^2V' \\ &\equiv \Gamma\Lambda\Gamma'\end{aligned}$$

- v_i: eigenvectors of $X'X$
- d_i^2: eigenvalues of $X'X$
- u_i: eigenvectors of XX'

$$XX' = UD^2U'$$

$X'X$ and XX' have all the same non-zero eigenvalues.

Homework (A3.3')**.**

(i) $C_{row} = UD$*: gives coordinates of* X*'s rows in the orthogonal basis* $(v_1, \ldots, v_r)$*.*

$$\textit{ith row of } X = \sum_{i=1}^{r} a_i v_i$$

$a_1, \ldots, a_r$ *is the* i*th row of* C_{row}

(ii) $C_{col} = DV'$*: gives coordinates of* X*'s columns in the orthogonal basis* $(u_1, \ldots, u_r)$*.*

$$\textit{ith row of } X = \sum_{i=1}^{r} a_i v_i$$

$a_1, \ldots, a_r$ *is the* i*th row of* C_{row}

Proof. Write $A = X'X = VD^2V'$. Let $U = XVD^{-1}$.

(i)

$$\begin{aligned}U'U &= D^{-1}V'X'XVD^{-1} \\ &= D^{-1}V'(VD^2V')VD^{-1} \\ &= I_r\end{aligned}$$

(ii)

$$\begin{aligned}UDV' &= XVD^{-1}DV' \\ &= X\end{aligned}$$

□

Homework (A3.4)**.** *Show*

(i) $L_{col}(X) = L_{col}(U)$
(ii) $L_{row}(X) = L_{col}(V)$

A3.4 Pseudo inverse. Regular inverse, $A = \Gamma\Lambda\Gamma'$ full rank ($\text{rank}(A) = p$) means

$$A^{-1} = \Gamma\Lambda^{-1}\Gamma'$$

In general, X is $n \times p$, with $\text{rank}(X) < p$, $p \leq n$ The pseudo inverse has $\text{rank}(A) < p$. Define a pseudo inverse (note that it is not unique) to be

$$X^- = \underbrace{V}_{p\times r}\underbrace{D^{-1}}_{r\times r}\underbrace{U'}_{r\times n}$$

so X^- is $p \times n$.
Facts:

(a) $L_{col}(X^-) = L_{col}(V) = L_{col}(X)$
(b) $L_{row}(X^-) = L_{col}(U) = L_{col}(X)$
(c)

$$\begin{aligned} XX^- &= UDV'VD^{-1}U' \\ &= UU' \\ &= \text{“}P_{col}\text{”} \end{aligned}$$

This is a projection from $\mathbb{R}^n$ to $L_{col}(X)$.
d)

$$X^-X = VV' = \text{“}P_{row}\text{”}$$

(e)

$$XX^-X' = X$$

(One version of definition of pseudo inverse)

HWA3.5 Solving linear equation Let

$$\underbrace{\eta}_{n\times 1} = \underbrace{X}_{n\times p}\underbrace{\beta}_{p\times 1}$$

assuming $\eta \in L_{col}(X)$.

- $\beta_0 = X^-\eta \in L_{col}(X^-) \equiv L_{col}(X)$ and $X\beta_0 = XX^-\eta = \eta$.
- $\beta = \beta_0 + \beta^\perp$, $\beta^\perp \perp L_{row}(X)$ and $\beta_0 \in L_{row}(X)$, then

$$\begin{aligned} X\beta &= X\beta_0 + X\beta^\perp \\ &= X\beta_0 \\ &= \eta \end{aligned}$$

- $\beta_0 \perp \beta^\perp$
- $\|\beta\|^2 = \|\beta_0\|^2 + \|\beta^\perp\|^2 > \|\beta_0\|^2$

so we have that

- β_0 is a solution
- It is the only one that lives in $L_{row}(X)$
- It is the “shortest” among all solutions

In general, $\eta \notin L_{col}(X)$.

$$\begin{aligned} \hat{\beta} &= X^-\eta \\ &= P_{col}(\eta) \\ &= \text{“}\hat{\eta}\text{”} \\ &= \text{least square estimate} \end{aligned}$$

4. September 16th, 2013

4.1. **Pseudo-Inverse.** If X is $n \times p$, $p \leq n$ with X full rank and $r = p \leq n$, with $X = UDV'$, a pseudo-inverse is defined as

$$X^- = VD^{-1}U'$$

- $L_{row}(X^-) = L_{col}(X)$
- $X^- = VD^{-1}U' \cdot UDV' = I_p$

Now, if we write

$$X = \begin{bmatrix} \vdots & \vdots & \vdots \\ x_1 & \dots & x_p \\ \vdots & \vdots & \vdots \end{bmatrix} \qquad X^- = \begin{bmatrix} \dots & x_1^- & \dots \\ & \vdots & \\ \dots & x_p^- & \dots \end{bmatrix}$$

where

$$(x_i^-, x_j) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & \text{otherwise} \end{cases}$$

Example 3 (p=2)**.** *Let*

$$x_1^\perp = x_1 - \frac{(x_1, x_2)}{(x_2, x_2)} x_2 \quad \textit{part of } x_1 \perp \textit{ to } x_2$$
$$x_2^\perp = x_2 - \frac{(x_1, x_2)}{(x_1, x_1)} x_1 \quad \textit{part of } x_2 \perp \textit{ to } x_1$$

We also have that $(x_1^-)' x_2 = 0$*, so*

$$x_1^- \propto x_1^\perp$$

meaning that $x_1^- = c x_1^\perp$ *for some* c*. We also have that* $(x_1^-)' x_1 = 1$ *and combining these two, we get that*

$$c(x_1^\perp)' x_1 = 1$$

so

$$c\|x_1^\perp\|^2 \Rightarrow c = \frac{1}{\|x_1^\perp\|^2}$$

Multiplying both sides by $x_1^\perp$*, we get*

$$x_1^- = \frac{x_1^\perp}{\|x_1^\perp\|^2} \Rightarrow \|x_1^-\| = \frac{1}{\|x_1^\perp\|}$$

Homework (A3.5)**.** *In general, for* $X_{n \cdot p}$ *full rank with* $p \neq n$*, show*

(i)

$$x_j^- = \frac{x_j^\perp}{\|x_j^\perp\|^2}$$

, *where* $x_j^\perp$ *is part of* x_j *not in* $L(x_1, \dots, x_{j-1}, x_{j+1}, \dots x_p)$

(ii) *(This part should follow trivially)*

$$\|x_j^-\| \|x_j^\perp\| = 1$$

A4.1 Orthogonal Representation. X is $n \times p$ and $\text{rank}(X) = r$

$$X = UDV'$$
$$= \sum_{k=1}^{r} d_k U_k V_k'$$

with U_k being the kth column of U and V_k being the kth column of V.

$$U = \begin{bmatrix} \vdots & \vdots & \vdots \\ U_1 & \dots & U_r \\ \vdots & \vdots & \vdots \end{bmatrix}$$

$$V = \begin{bmatrix} \vdots & \vdots & \vdots \\ V_1 & \dots & V_r \\ \vdots & \vdots & \vdots \end{bmatrix}$$

Think of $\underbrace{U_k V_k'}_{n \times p}$ as a vector in $\mathbb{R}^{np}$, by stacking it. Then $\underset{\sim}{b_k} \in \mathbb{R}^{np}$. Think of X as a vector in $\mathbb{R}^{np}$ as well. Write it as $\underset{\sim}{X}$

$$\underset{\sim}{X} = \sum_{k=1}^{r} d_k \underset{\sim}{b_k}$$

We have

$$\|X\|^2_{Frobenius} = \|\underset{\sim}{X}\|^2$$

Homework (A4.1)**.** *Show that* $\underset{\sim}{b_k}$ *are orthogonal vectors with unit length in* $\mathbb{R}^{np}$*.*

4.2. **Idea of Principle Component Analysis (PCA).**

$$d_1 \geq d_2 \geq \dots d_r > 0$$

We hope that we have K, presumably $K < r$ such that

$$\hat{X}(K) = \sum_{k=1}^{K} d_k U_k V_k' \approx X$$

$$\begin{aligned} \|X - \hat{X}(K)\|_F^2 &= \|\underset{\sim}{X} - \underset{\sim}{\hat{X}}(K)\|^2 \\ &= \left\| \sum_{k=K+1}^{r} d_k \underset{\sim}{b_k} \right\|^2 \\ &= \sum_{k=K+1}^{r} d_k^2 \end{aligned}$$

We hope that

$$\frac{\sum_{k=K+1}^{r} d_k^2}{\sum_{k=1}^{r} d_k^2}$$

is small or

$$R^2 \equiv \frac{\sum_{k=1}^{K} d_k^2}{\sum_{k=1}^{r} d_k^2}$$

is close to 1

Homework (A4.2)**.** *Show that*

$$\|\underset{\sim}{X} - \underset{\sim}{\hat{X}}(K)\|^2 = \sum_{k=K+1}^{r} d_k^2$$

Example 4. *Score txt with a matrix of size* 88×5

- *88 students in a class*
- *test scores for 5 exams*

"Two" major features

- *quality of the student*
- *closed/open book*

Let us first let

- $X = [x_1, x_2, \dots, x_5]$
- Normalize each column with mean 0 and sd 1
- $X = UDV'$ (rank is usually 5 for real data) In the handout,

$$d = 16.64, 8.02, 6.22, 5.80, 4.63$$

$$R^2 = \frac{16.64^2 + 8.02^2}{400} \approx 80\%$$

Let $v = [v_1, v_2, \dots v_5]$.

$$\begin{aligned} X &= UDV' \\ XV &= UD \\ d_1 U_1 &= XV_1 \\ d_2 U_2 &= XV_2 \end{aligned}$$

What this means is that

$$U_1 \propto XV_1$$

the solid line corresponds to v_1 because the line is roughly flat, we take this to be constant. We can assume wlog v_1 is a vector of ones because it only differs by a constant. This means that $U_1 \propto XV_1$ is the average of a student's scores. (Rows of X are the students. Columns are the scores).

$$\begin{aligned} \hat{X}(1) &= d_1 U_1 V_1' \\ &= d_1 [0.4U_1, 0.43U_1, 0.5U_1, 0.457U_1, 0.438U_1] \\ \hat{X}(2) &= [0.4 d_1 U_1 + 0.646 d_2 U_2, \dots] \end{aligned}$$

The first principle component is the x-axis and the second one is the y-axis.

Homework (A4.3)**.**

(a) *Use the R function "svd" to compute* $\hat{X}(1)$ *and* $\hat{X}(2)$ *for the "score" data. Plot* $\hat{\underset{\sim}{X}}(1)$ *versus* $\underset{\sim}{X}$ *and similarly* $\hat{\underset{\sim}{X}}(2)$ *versus* $\underset{\sim}{X}$*.*

(b) *Compute* $\|\underset{\sim}{X} - \hat{\underset{\sim}{X}}(K)\|^2$ *for* $K = 1, 2$ *and verify numerically the relationship*

$$R^2 = \frac{\sum_{k=1}^{K} d_k^2}{\sum_{k=1}^{r} d_k^2} = 1 - \frac{\|\underset{\sim}{X} - \hat{\underset{\sim}{X}}(K)\|^2}{\|\underset{\sim}{X}\|^2}$$

4.3. **Partitioned Inverse.** For G nonsingular,

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \qquad H \equiv G^{-1} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}$$

where

$$H = \begin{bmatrix} (G_{11} - G_{12}G_{22}^{-1}G_{21})^{-1} & -G_{11}^{-1}G_{12}(G_{22} - G_{21}G_{11}^{-1}G_{12})^{-1} \\ -G_{22}^{-1}G_{21}(G_{11} - G_{12}G_{22}^{-1}G_{21})^{-1} & (G_{22} - G_{21}G_{11}^{-1}G_{12})^{-1} \end{bmatrix}$$

Theorem 2 (Woodbury's Theorem)**.** *We have* A*, which is* $p \times p$ *and* B *which is* $q \times q$ *and both are non-singular. Then (where* U *is* $p \times q$ *and* V *is* $q \times p$*)*

$$[A + UBV]^{-1} = A^{-1} - A^{-1}U\left[B^{-1} + VA^{-1}U\right]^{-1}VA^{-1}$$

Example 5 (q=1)**.** *So we have* b *is a scalar*

$$(A + buu')^{-1} = A^{-1}\left[A - \frac{uu'}{\frac{1}{b} + u'A^{-1}u}\right]A^{-1}$$

Especially when $A = I_p$*, this implies*

$$(I + buu')^{-1} = \left(I - \frac{uu'}{\frac{1}{b} + \|u\|^2}\right)$$

Homework (A4.4)**.** *Apply this to*

$$\Sigma = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & 1 & \rho \\ \rho & \dots & \rho & 1 \end{bmatrix}$$

For what value of ρ *is* $\Sigma > 0$*?*

For the homework above,

$$(G^{-1})_{22} = (G_{22} - G_{21}G_{11}^{-1}G_{12})^{-1}$$

Now, suppose G is a Gram matrix, where $\text{rank}(G) = p$

$$\underbrace{G}_{p\times p} = \underbrace{X'}_{p\times n}\underbrace{X}_{n\times p} \qquad X = [X_1, X_2] \qquad G_{22} = X_2'X_2$$

Now, X_2 can be spanned by $\hat{X_2}, \overset{\perp}{X}_2$

$$\hat{X_2} = P_{col}(X_1)X_2$$
$$\overset{\perp}{X}_2 = P_{col}^{\perp}(X_1)X_2$$

Now, use note A2.5, class 2 and

$$G_{22} - G_{21}G_{11}^{-1}G_{12} = (\overset{\perp}{X}_2)'\overset{\perp}{X}_2 \equiv \overset{\perp}{G}_{22}$$

Now, $(G^{-1})_{22} = \overset{\perp}{G}_{22}$

Homework (A4.5)**.** *Show that*

$$G^{-1} = X^{-}(X^{-})'$$

and relate this to

$$\|X_j^{-}\|\|X_j^{\perp}\| = 1$$

4.4. **Statistical Interpretation.**

$$U = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} \qquad \mathrm{Cov}(U) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where U_1 is p_1 and U_2 is p_2. Then

$$(\Sigma^{-1})_{22} = (\overset{\perp}{\Sigma}_{22})^{-1} = (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}$$

where

$$\overset{\perp}{\Sigma}_{22} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

is the covariance matrix of

$$U_2^{\perp} = U_2 - \Sigma_{21}\Sigma_{11}^{-1}U_1$$

$$GRAPHDIAGRAM$$

$$\langle I_2^{\perp}, U_2 \rangle$$

4.5. **Partial Correlation.**

Definition 4 (Partial Correlation)**.** *If $\Sigma_{ij} = 0$, we call i, j uncorrelated. If $(\Sigma^{-1})_{ij} = 0$, then U_i and U_j has 0 partial correlation (correlation, after (U_i, U_j) are regressed over all other coordinates.)*

Homework (A4.5)**.** *Verify the last statement of the above definition.*

5. September 23rd, 2013

5.1. **B part: Multivariate normal distribution.**

Definitions and first results.

Definition 5 (Standard Normal)**.** *We have*

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_p \end{pmatrix} \sim \mathcal{N}_p(\mathbf{0}, I_p)$$

if $Z_i \overset{iid}{\sim} \mathcal{N}(0,1)$

$$f_Z(z) = \frac{1}{(2\pi)^{p/2}} e^{-\|z\|^2/2}$$

Definition 6 (Characteristic Function)**.**

$$\begin{aligned} \mathbb{E}\left[e^{a'Z}\right] &= \frac{1}{(2\pi)^{p/2} e^{-\frac{1}{2}\|z\|^2 + a'z}} \\ &= e^{\frac{1}{2}\|a\|^2} \\ \mathbb{E}\left[e^{it'Z}\right] &= e^{-\frac{1}{2}\|t\|^2} \end{aligned}$$

Definition 7 (General Normal)**.** *Let $\mu \in \mathbb{R}^p$, $\Sigma^{1/2}$ is $p \times p$. Let*

$$\begin{aligned} X &= \mu + \Sigma^{1/2} Z \\ &\sim \mathcal{N}_p(\mu, \Sigma) \end{aligned}$$

where

$$\Sigma = (\Sigma^{1/2})(\Sigma^{1/2})' \geq 0$$

When $\Sigma^{1/2}$ is full rank, $\Sigma > 0$, the density has a simple formula. When $\Sigma^{1/2}$ is not full rank,

$$\Sigma = \Gamma \begin{bmatrix} d_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & d_r & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix} \Gamma'$$

where each matrix is $p \times p$, which equals

$$\Sigma = \Gamma D (\Gamma D)'$$

and we have

$$X = \mu + \Gamma \begin{pmatrix} d_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & d_r & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{pmatrix} Z$$
$$= \mu + \Gamma_1 \begin{pmatrix} d_1 z_1 \\ \vdots \\ d_r z_r \end{pmatrix}$$

where

$$\Gamma = \begin{bmatrix} \Gamma_1 & \Gamma_2 \end{bmatrix}$$

Γ_1 *is* $p \times r$ *and* Γ_2 *is* $p \times (p-r)$.

Facts:

(1) $\mathbb{E}[X] = \mathbb{E}\left[\mu + \Sigma^{1/2} Z\right] = \mu$ and $\mathrm{Var}(X) = \mathrm{Var}(\mu + \Sigma^{1/2} Z)$

(2) If $\mathrm{rank}(\Sigma^{1/2}) =$ and

$$f_X(x) = \frac{(2\pi)^{-\rho/2}}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}$$

then

$$f_Z(z) = \frac{1}{(2\pi)^{\rho/2}} e^{-\frac{1}{2}\|z\|^2}$$

and

$$\frac{\mathrm{d}z}{\mathrm{d}x} = (\Sigma^{1/2})^{-1}$$

So we have

$$f_X(x) = f_Z(z)|_{z=(\Sigma^{1/2})^{-1}(x-\mu)} \left|\frac{\mathrm{d}z}{\mathrm{d}x}\right|$$
$$= \frac{1}{(2\pi)^{\rho/2}} e^{-\frac{1}{2}\|z\|^2}\Big|_{z=\Sigma^{-1/2}(x-\mu)} \frac{1}{|\Sigma|^{1/2}}$$

(3) $\varphi(t) = \psi(t)$ for t in an interval containing 0, which implies that the CDFs are the same

$$\varphi_X(t) = \mathbb{E}\left[e^{it'X}\right]$$
$$= \mathbb{E}\left[e^{it(\mu+\Sigma^{1/2}Z)}\right]$$
$$= \mathbb{E}\left[e^{it'\mu+(it'\Sigma^{1/2})'Z}\right]$$
$$= e^{it'u-\frac{1}{2}\|\Sigma^{1/2}t\|^2}$$
$$= e^{it'\mu-\frac{1}{2}t'\Sigma t}$$

(4) For any choice of (μ, Σ) such that $\Sigma \geq 0$, there is a unique $\mathcal{N}_p(\mu, \Sigma)$

Proof. The characteristic functions equals

$$e^{it'\mu-\frac{1}{2}t'\Sigma t}$$

□

(5) $X \in \mu + L_{col}(\Sigma)$ with probability 1.

Proof.

$$X = \mu + \Sigma^{1/2} Z \in \mu + L_{col}(\Sigma^{1/2}) \equiv \mu + L_{col}(\Sigma)$$

□

(6) $Y = v + \Delta X \sim \mathcal{N}_q(v + \Delta\mu, \Delta\Sigma\Delta')$ where v has dimension q, Δ is $q \times p$ and X has dimension p.

Proof. Again, we do this by characteristic functions. We only need to show that the characteristic function of the left equals the characteristic function on the right. □

(7) "Kurtosis"

$$\mathbb{E}[(x_1-\mu_1)(x_2-\mu_2)(x_3-\mu_3)(x_4-\mu_4)] = \sigma_{12}\sigma_{34}+\sigma_{13}\sigma_{24}+\sigma_{14}\sigma_{23}$$

$$\begin{aligned}\mathbb{E}[X] &\text{ mean}\\ \mathbb{E}\big[(X-\mu)^2\big] &\text{ variance}\\ \mathbb{E}\big[(X-\mu)^3\big] &\text{ skewness}\\ \mathbb{E}\big[(X-\mu)^4\big] &\text{ kurtosis}\end{aligned}$$

$$K(X) = \frac{\mathbb{E}\big[(X-\mu)^4\big]}{(\mathbb{E}[(X-\mu)^2])^2} = \frac{3}{1^2} = 3$$

Let $g(t) = \log(\mathbb{E}\big[e^{tX}\big])$, which is the moment generating function. In the univariate case,

$$g(t) = \sum_{k=1}^{\infty} \frac{1}{n!} K_n t^n$$

(8) If we have $\mu = 0$, $\Sigma^{1/2} \equiv \Gamma$ which is $n \times n$ and orthogonal, and

$$X = \Gamma Z \sim \mathcal{N}(\mathbf{0}, \Gamma\Gamma') \sim \mathcal{N}(\mathbf{0}, I_p)$$

then this implies that the Standard Normal is spherical and symmetric.

(9) If $\mu = 0$, $\Sigma^{1/2} = \begin{bmatrix}\Sigma & \mathbf{0}\end{bmatrix}$ where Γ is $p \times n$ and $\mathbf{0}$ is $p \times (p-n)$ and

$$X = \Gamma Z \sim \mathcal{N}(\mathbf{0}, \underbrace{\Gamma\Gamma'}_{\text{projection matrix}})$$

then

$$Y \equiv \Gamma' X \sim \mathcal{N}_n(\mathbf{0}, I_n)$$

"The spherical normal projects into a lower dimensional normal"

(10) **Very important/useful** - "Haar Measure"
The unit-norm vector

$$U = \frac{Z}{\|Z\|}$$

is uniformly distributed on the surface of the unit sphere of $\mathbb{R}^p$, independently of $\|Z\|^2 \sim \chi^2_p(0)$.

(11) If $\begin{pmatrix}X\\Y\end{pmatrix} \sim \mathcal{N}_{p+q}\left(\begin{bmatrix}\mu\\\nu\end{bmatrix}, \begin{bmatrix}\Sigma_{XX} & \Sigma_{XY}\\ \Sigma_{YX} & \Sigma_{YY}\end{bmatrix}\right)$ then

$$X \sim \mathcal{N}_p(\mu, \Sigma_{XX}) \qquad \varphi_X(t) = \varphi_{X,Y}(t,0)$$

(12) If $\Sigma_{X,Y} = \underbrace{\mathbf{0}}_{p\times q}$, then $X \perp\!\!\!\perp Y$. The easiest way to show this is by using characteristic functions.

$$\varphi_{X,Y}(t,s) = \varphi_X(y)\varphi_Y(t)$$

(13)

$$(Y|X) \sim \mathcal{N}_q(\nu + \Sigma_{YX}\Sigma_{XX}^{-1}(x-\mu), \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY})$$

We can let

$$Y = \hat{Y} + \overset{\perp}{Y}$$

We have that

$$\hat{Y} = \nu + \Sigma_{YX}\Sigma_{XX}^{-1}(x-\mu)$$
$$\mathrm{Cov}(\overset{\perp}{Y}) = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$

(14) $\mathrm{Cov}(Y|X) = \mathrm{Var}(\overset{\perp}{Y})$. This is the same as saying that the residual covariance after the projection is equal to the conditional covariance.

Homework (B1.1)**.** *The unit-norm vector*

$$U = \frac{Z}{\|Z\|}$$

is uniformly distributed on the surface of the unit sphere of $\mathbb{R}^p$*, independently of* $\|Z\|^2 \sim \chi^2_p(0)$*.*

(a) Use multivariate polar coordinates to show independence
(b) How is the uniform density on the surface of the sphere in $\mathbb{R}^p$ *expressed in terms of* $\theta_1, \theta_2, \ldots, \theta_p$*?*

Hint: Let

$$\begin{aligned} Z_1 &= r\cos\theta_1 \\ Z_2 &= r\sin\theta_1\cos\theta_2 \\ &\vdots \\ Z_p &= r\sin\theta_1\dots\sin\theta_{p-1} \end{aligned}$$

We should get something that looks like

$$f_Z(z) = g(r)h(\theta_1,\dots,\theta_p)$$

Homework (B1.2)**.** *Show that if* $\Sigma_{X,Y} = \underbrace{\mathbf{0}}_{p\times q}$*, then* $X \perp\!\!\!\perp Y$*, then*

$$\varphi_{X,Y}(t,s) = \varphi_X(t)\varphi_Y(s)$$

Homework (B1.3 - Multivariate Cauchy)**.** $X \sim \mathcal{N}_p(\mathbf{0},\Sigma)$*,* $Z_i \sim \mathcal{N}(0,1)$*,* $Z_0 \perp\!\!\!\perp X$*. Now* $Y = \frac{X}{Z_0}$ *has characteristic function*

$$\begin{aligned} \varphi_Y(t) &= e^{-(t'\Sigma t)^{1/2}} \\ &= e^{-|t|} \end{aligned}$$

and has density

$$f_Y(y) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\pi^{\frac{p+1}{2}}|\Sigma|^{1/2}}(1+y'\Sigma^{-1}y)^{-\frac{p+1}{2}}$$

In the univariate case, a Cauchy distribution has density

$$\frac{1}{\pi}(1+y^2)^{-1} \to e^{-|t|}$$

Its characteristic function, the double exponential??

$$\frac{1}{2}e^{-|t|} \to \frac{1}{1+y^2}$$

$$f(t) = \frac{1}{2\pi}\int \varphi(t)e^{-itX}\,\mathrm{d}t$$

5.1.1. *Repeated Sampling.* Let $x_1, x_2, \dots, x_n \overset{iid}{\sim} \mathcal{N}_p(\mu,\Sigma)$, $\Sigma > 0$. Write

$$X = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}$$

which is $p \times n$. Typically, we would write

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix}$$

which is $n \times p$. Now, the density is

$$f_{\mu,\Sigma}(x_1,\dots,x_n) = (2\pi)^{-np/2}|\Sigma|^{-n/2}e^{-\frac{1}{2}\sum_{i=1}^n (x_i-\mu)'\Sigma^{-1}(x_i-\mu)}$$

Idea:

$$\begin{aligned} &\sum_{i=1}^n (x_i-\mu)'\Sigma^{-1}(x_i-\mu) \\ &= \sum_{i=1}^n \operatorname{tr}\left(\Sigma^{-1}(x_i-\mu)(x_i-\mu)'\right) \\ &= \operatorname{tr}\left[\Sigma^{-1}\sum_{i=1}^n (x_i-\mu)(x_i-\mu)'\right] \\ &= \operatorname{tr}\left[\Sigma^{-1}\sum_{i=1}^n (x_i-\bar{x})(x_i-\bar{x})' + n(\bar{x}-\mu)(\bar{x}-\mu)'\right] \\ &= \operatorname{tr}\left(\Sigma^{-1}(n-1)S_n\right) + \operatorname{tr}\left(n\Sigma^{-1}(\bar{x}-\mu)(\bar{x}-\mu)'\right) \\ &= \operatorname{tr}\left(\Sigma^{-1}(n-1)S_n\right) + n(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu) \\ &= (n-1)\operatorname{tr}\left(\Sigma^{-1}S_n\right) + n(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu) \end{aligned}$$

There is a famous algorithm called graphic lasso or glasso. The goal of this algorithm is to estimate Σ^{-1} assuming it is sparse.

5.1.2. *Kronecker-product and random matrices.* Suppose you have independent sampling

$$\underbrace{\underset{\sim}{X}}_{p\times n} = (X_1, X_2, \dots X_n) \qquad X_i \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$$

$$\underbrace{\overset{\circ}{X}}_{np\times 1} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

$$\overset{\circ}{X} \sim \mathcal{N}(\overset{\circ}{\mu}, \overset{\circ}{\Sigma})$$

where $\overset{\circ}{\mu}$ is $np \times 1$ and $\overset{\circ}{\Sigma}$ is $np \times np$.

$$\overset{\circ}{\mu} = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} \qquad \overset{\circ}{\Sigma} = \operatorname{diag}(\Sigma, \dots, \Sigma)$$

Consider the Linear Transformation

$$\underbrace{A}_{a\times p}\underbrace{X}_{p\times n}\underbrace{B}_{n\times b}$$

where A, B are nonstochastic matrices.

$$Ax_i \sim \mathcal{N}(\mathbf{0}, A\Sigma A')$$
$$Ax_j \sim \mathcal{N}(\mathbf{0}, B'IB)$$

6. September 25th, 2013

6.1. **Wishart Distribution.** We have

$$X = \begin{bmatrix} X_1 & \dots & X_n \end{bmatrix}_{p\times n} \qquad X_i \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$$

The Wishart is XX'. What is its density?

$$X_{p\times n} = UDV' \qquad X_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1)$$

with prob. $\operatorname{rank}(D) = p$

- U: Uniform distribution over all $n \times p$ matrices with unit-norm, orthogonal columns
- V is the same thing but for a $p \times p$ matrix
- D is a diagonal matrix, $\sqrt{X^2}$
- U, D, V are independent

B2.1 Kronecker Product. **Motivation:**

$$Y = AXB$$

Stringing should work, but it's very clumsy.

Definition 8 (Kronecker-product)**.**

$$\underbrace{A}_{p\times q} \bigotimes \underbrace{B}_{r\times s} = \textit{matrix of size } pr \times qs$$
$$= \begin{bmatrix} b_{11}A & b_{12}A & \dots & b_{1s}A \\ \dots & \dots & \dots & \dots \\ b_{r1}A & b_{r2}A & \dots & b_{rs}A \end{bmatrix}_{pr\times qs}$$

Warning: *In R, the "Kronecker" is different where A and B are reversed.*

Properties:

- $(A \bigotimes B)' = A' \bigotimes B'$
- $(A \bigotimes B)^{-1} = A^{-1} \bigotimes B^{-1}$ if A^{-1}, B^{-1} exist
- $(A \bigotimes B)(C \bigotimes D) = (AC) \bigotimes (BD)$

- If we have $A_{p\times p}, B_{r\times r}$ and $Ax = \lambda x$, $By = \mu y$ and

$$x \bigotimes y = \begin{pmatrix} y_1 x \\ y_2 x \\ \vdots \\ y_r x \end{pmatrix}$$

then

$$(A \bigotimes B)(x \bigotimes y) = \lambda\mu x \bigotimes y = (Ax) \bigotimes (By)$$

- $\overset{\circ}{Y} \equiv A\overset{\circ}{X}B = (A \bigotimes B')\overset{\circ}{X}$

Homework (B2.1). *Verify* $\overset{\circ}{Y} \equiv (A\overset{\circ}{X}B) = (A \bigotimes B')\overset{\circ}{X}$ *and* $(A \bigotimes B)(C \bigotimes D) = (AC) \bigotimes (BD)$

6.2. **Normal Matrix Distribution.** Suppose $\overset{\circ}{X}$, which is $np \times 1$ is obtained (by stringing) from

$$\underset{\sim}{X}_{p\times n} = (x_1, x_2, \ldots, x_n)$$

which has a joint normal distribution such that there are two non-random matrices $\Sigma_{p\times p}$ and $\Delta_{n\times n}$.

- $\mathbb{E}[X_{ij}] = \mu_{ij}$
- $\mathrm{Cov}(X_{i_1,j_1}, X_{i_2,j_2}) = \Delta_{j_1 j_2}\sigma_{i_1 i_2}$ where $1 \le i_1, i_2 \le p$, $1 \le j_1, j_2 \le n$.

Then

$$\underbrace{\mathrm{Cov}(\overset{\circ}{X})}_{np\times np} = \Sigma \bigotimes \Delta$$

Now, in light of this, we introduce a new notation, for matrices.
Let

$$X_{p\times n} \sim \mathcal{N}_{p\times n}(\mu, \Sigma \bigotimes \Delta)$$

with

$$\mu = (\mu_1, \mu_2, \ldots, \mu_n) \equiv (\mu_{ij})_{\substack{1\le i\le p \\ 1\le j\le n}}$$

which is $p \times n$. Most of the time, we no longer need stringing. This implies different rows have common covariance structure up to constant time. Similarly, different columns have common covariance structure up to common time.

What are Σ and Δ?

- $\mathrm{Cov}(X_{j_1}, X_{j_2}) = \Delta_{j_1 j_2}\Sigma$ which is the column covariance.
- Rewrite

$$X = \begin{bmatrix} \cdots & v_1' & \cdots \\ \cdots & v_2' & \cdots \\ \cdots & \vdots & \cdots \\ \cdots & v_p' & \cdots \end{bmatrix}$$

and

$$\mathrm{Cov}(v_{i_1}, v_{i_2}) = \sigma_{i_1 i_2}\Delta$$

which is the row covariance.

Facts:

(1) $X_i \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma) \Leftrightarrow \underset{\sim}{X} \sim \mathcal{N}_{p\times n}\big(\underbrace{\underset{\sim}{\mu}(1,1,\ldots,1)}_{p\times n}, \Sigma \bigotimes I_n\big)$

(2) $\underset{\sim}{X}' \sim \mathcal{N}_{n\times p}(\mu', \Delta \bigotimes \Sigma)$, if $\underset{\sim}{X} \sim \mathcal{N}_{p\times n}(\mu, \Sigma \bigotimes \Delta)$

(3) If Σ and Δ have spectral representation:

$$\Sigma = \underbrace{W_1}_{p\times k_1}\underbrace{D_1}_{k_1\times k_1}\underbrace{W_1'}_{k_1\times p} \qquad \Delta = \underbrace{W_2}_{n\times k_2}\underbrace{D_2}_{k_2\times k_2}\underbrace{W_2'}_{k_2\times n}$$

then

$$\underset{\sim}{X} = \underset{\sim}{\mu} + W_1 D_1^{1/2} Z D_2^{1/2} W_2' \equiv \underset{\sim}{\mu} + \Sigma^{1/2} Z \Delta^{1/2}$$

where $Z \sim \mathcal{N}_{k_1\times k_2}(\mathbf{0}, I_{k_1} \bigotimes I_{k_2})$, where equivalently $Z_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$. We call Z the "standard normal matrix".

Proof.

$$Z = D_1^{-1/2} W_1' (\underset{\sim}{X} - \underset{\sim}{\mu}) W_2 D_2^{-1/2}$$

□

(4)

Theorem 3. *If*

$$\underset{\sim}{X} \sim \mathcal{N}_{p\times n}(\underset{\sim}{\mu}, \Sigma \bigotimes \Delta)$$

and

$$\underbrace{\underset{\sim}{Y}}_{a\times b} = \underbrace{A}_{a\times p}\underbrace{\underset{\sim}{X}}_{p\times n}\underbrace{B}_{n\times b}$$

then

$$\underset{\sim}{Y} \sim \mathcal{N}_{a\times b}(A\mu B, (A\Sigma A') \bigotimes (B'\Delta B))$$

Proof.

$$\underset{\sim}{X} = \underset{\sim}{\mu} + \Sigma^{1/2} Z (\Delta^{1/2})'$$

and since $\mathcal{N}_{p\times n}(\underset{\sim}{\mu}, \Sigma \bigotimes \Delta)$

$$A\underset{\sim}{X}B = A\underset{\sim}{\mu}B + A\Sigma^{1/2} Z (B'\Delta^{1/2})$$

since $\mathcal{N}_{a\times b}(A\underset{\sim}{\mu}B, A\Sigma A' \bigotimes B'\Delta B)$ □

Homework (B2.2)**.** *We observe $X_i \overset{iid}{\sim} \mathcal{N}_p(\mu, \Sigma)$, $i = 1, 2, \ldots, n$ with Σ known to have unit diagonals ($\sigma_{ii} = 1$), and compute*

$$Z_i = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} X_{ij} \qquad i = 1, 2, \ldots, p$$

to test the null hypothesis $\mu_i = 0$ What is the correlation matrix of

$$\underbrace{Z}_{p\times 1} = \begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix}$$

B2.2 Rotational Invariance. Write $\underset{\sim}{X} \sim \mathcal{N}_{p\times n}(\mathbf{0}, \Sigma \bigotimes I_n)$ and write

$$\underset{\sim}{X} = \begin{pmatrix} \ldots & v_1' & \ldots \\ \ldots & v_2' & \ldots \\ & \vdots & \\ \ldots & v_p' & \ldots \end{pmatrix}$$

Homework (B2.3)**.** $v_i \sim \mathcal{N}_n(\mathbf{0}, \sigma_{ii} I_n)$

$v_1, \ldots, v_p$ may be correlated. It turns out $v_1, \ldots, v_p$ together have spherical symmetry.

Corollary. *For any $n \times n$ orthogonal matrix $\Gamma_{n\times n}$*

$$\underbrace{\Gamma(v_1, v_2, \ldots, v_p)}_{new\ matrix} \equiv \Gamma X' \sim X'$$

Proof.

$$\begin{aligned} \Gamma X' &\sim \mathcal{N}_{p\times n}(\mathbf{0}, \Sigma \bigotimes \Gamma'\Gamma) \\ &\sim \mathcal{N}_{p\times n}(\mathbf{0}, \Sigma \bigotimes I_n) \end{aligned}$$

X' is an $n \times p$ matrix. The column space is

$$L_{col}(v_1, \ldots, v_p) \equiv L_{row}(X)$$

$\Gamma X'$ is an $n \times p$ matrix. The "random subspace" is defined by $L_{col}(v_1, \ldots, v_p)$. This "random subspace" is uniform over all p-dimensional subspaces of $\mathbb{R}^n$.

Let $Q_{n\times p}$ be all $n \times p$ matrices such that $Q'Q = I_p$.

$$X = UDV'$$

where X is $n \times p$ and U is $n \times p$ with orthogonal columns. □

Example 6 (Removing Means)**.** *Let $\bar{p} = I_n - \mathbf{1}\mathbf{1}'/n$. Starting with*

$$\underbrace{\underset{\sim}{X}}_{p\times n} \sim \mathcal{N}_{p\times n}(\mu \mathbf{1}_n', \Sigma \bigotimes I_n)$$

and

$$\underset{\sim}{Y} = \underbrace{\underset{\sim}{X}}_{p\times n}\underbrace{\bar{p}}_{n\times n} = \left[X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X}\right]$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(1)

$$\underset{\sim}{Y} \sim \mathcal{N}_{p\times n}(\mathbf{0}, \Sigma \bigotimes \overset{\perp}{p})$$

$$\underset{\sim}{Y} \equiv \underset{\sim}{X}\overset{\perp}{p} \sim \mathcal{N}_{p\times n}(\underbrace{\mu \mathbf{1}_n' \overset{\perp}{p}}_{0}, \underbrace{\Sigma \bigotimes (\overset{\perp}{p})'\overset{\perp}{p}}_{\Sigma \bigotimes \overset{\perp}{p}})$$

(2) $\underset{\sim}{Y} \perp\!\!\!\perp \bar{X}$, $\bar{X} \sim \mathcal{N}_p(\mu, \Sigma/n)$

(3) $S = \underset{\sim}{Y}\underset{\sim}{Y}' = \sum_{j=1}^n (X_j - \bar{X})(X_j - \bar{X})' \perp\!\!\!\perp \bar{X}$

B2.3 Normal matrix Triangularization. Idea:

$$\underbrace{\underset{\sim}{X}}_{p\times n} = \begin{bmatrix} \cdots & v_1' & \cdots \\ \cdots & v_2' & \cdots \\ & \vdots & \\ \cdots & v_p' & \cdots \end{bmatrix} \sim \mathcal{N}_{p\times n}(\mathbf{0}, \Sigma \bigotimes I_n)$$

can be thought of as two parts.

- $L = L_{row}(\underset{\sim}{X}) = L_{col}(v_1, \ldots, v_p)$
- lengths/angles of $v_1, \ldots, v_p$.

Now,

- $L_{col}(v_1, \ldots, v_p)$ is sort of "uniform" in all p-dimensional subspaces.
- XX', which is $p \times p$ gives all lengths and angles of $v_1, \ldots, v_p$. This is the Wishart matrix.

Gram-Schmidt says that

$$\underbrace{V}_{n\times p} = \underbrace{W}_{n\times p}\underbrace{T}_{p\times p}$$

where T is an upper triangular matrix. The inner product matrix XX' is

$$V'V = T'W'WT = T'T$$

which is the product of a lower triangular matrix and an upper triangular matrix.

(1) If $\Sigma = I_p$, then $t_{ii}^2 \sim \chi^2_{n-i+1}$ and $t_{ij} \sim \mathcal{N}(0,1)$ for $i < j$.

Lemma. *The integral Jacobian from $\underset{\sim}{X}_{p\times n}$ to $T_{p\times p}$ is*

$$J(\underset{\sim}{X}) = c_1 \prod_{i=1}^p t_{ii}^{n-i}$$

where

$$c_1 = \frac{2^p \pi^{\frac{np}{2} - \frac{p^2}{4} + \frac{p}{4}}}{\prod_{j=1}^p \Gamma\left(\frac{n-j+1}{2}\right)}$$

Suppose that $\underset{\sim}{x} = m(x)$, where $m : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable, invertible, etc.

If $f(x)$ is the density of x, what is the density $\underset{\sim}{f}(\underset{\sim}{x})$?

$$f(\underset{\sim}{x}) = f(x) \cdot \left||M^{-1}(x)|\right|$$

Here, the Jacobian is

$$\left(\frac{\mathrm{d}x_i}{\mathrm{d}\underset{\sim}{x}_j}\right)_{n\times n}$$

7. September 30th, 2013

8. Jacobian and Integral Jacobian

Jacobian: Suppose $\tilde{x} = m(x)$ maps $\mathbb{R}^n \to \mathbb{R}^n$ in a continuously differentiable way such that $f(x)$ is the density of X. What is the density of $\tilde{X}$, say $\tilde{f}_{\tilde{X}}(\tilde{x})$

Answer:

$$M(x) = \left(\frac{\alpha \tilde{x}_i}{\alpha x_j}\right)_{1\le i,j\le n}$$

$$M^{-1}(x) = \left(\frac{\alpha x_i}{\alpha \tilde{x}_j}\right)_{1\le i,j\le n}$$

Then

$$\begin{aligned}\tilde{f}(\tilde{x}) &= f(x)\left||M^{-1}(x)|\right| \\ &= f(x)J(x \to \tilde{x})\end{aligned}$$

Example 7 (Multivariate Polar Coordinates)**.**

$$\begin{aligned}x_1 &= \rho\cos\theta_1 \\ x_2 &= \rho\sin\theta_1\cos\theta_2 \\ &\vdots \\ x_{n-1} &= \rho\sin\theta_1\sin\theta_2\ldots\sin\theta_{n-2}\cos\theta_{n-1} \\ x_n &= \rho\sin\theta_1\ldots\sin\theta_{n-2}\sin\theta_{n-1}\end{aligned}$$

where

$$\begin{aligned}&\rho \geq 0 \\ &0 \leq \theta_i \leq \pi \quad i = 1, 2\ldots n-2 \\ &0 \leq \theta_{n-1} \leq 2\pi\end{aligned}$$

Homework (B3.1)**.**

(a) *Show that*

$$\begin{aligned}&J(x \to \rho, \theta_1, \ldots, \theta_{n-1}) \\ &= \rho^{n-1}(\sin\theta_1)^{n-2}(\sin\theta_2)^{n-3}\cdots\sin(\theta_{n-2}) \\ &\equiv \rho^{n-1}\prod_{i=1}^{n-1}(\sin\theta_i)^{n-i-1}\end{aligned}$$

(b) *Show that the unit sphere in* $\mathbb{R}^n$ *has* $(n-1)$ *dimensional "area"*

$$A = \frac{2\pi^{n/2}}{\Gamma(n/2)}$$

Hint:

$$\int_0^{\pi/2}\sin^m(x)\,\mathrm{d}x = \frac{\sqrt{\pi}}{2}\frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2}+1)}$$

Integral Jacobian: Consider

$$\underbrace{X}_{n\times p} \longrightarrow \underbrace{S}_{\frac{p\times(p+1)}{2}} \equiv X'X$$

which is $\mathbb{R}^N \to \mathbb{R}^N$ Suppose $x \in \mathbb{R}^n$ maps into $y = (y_1, y_2) \in \mathbb{R}^n$, and say that $y = m(x)$. Suppose also that the density of X only depends on y_1.

$$f_X(x) = g(y_1)$$

What is the density of y_1?

Example 8. *If* $X \in \mathcal{N}_p(\mathbf{0}, I_p)$*, then*

$$\begin{aligned}f_X(x) &= \frac{1}{(2\pi)^{\rho/2}}e^{-\frac{1}{2}\|x\|^2} \\ &= \frac{1}{(2\pi)^{\rho/2}}e^{-\rho^2/2}\end{aligned}$$

and so

$$X \longrightarrow (\underbrace{\rho}_{y_1}, \underbrace{\theta_1, \ldots, \theta_{n-2}}_{y_2})$$

Now we wonder what is the density of ρ*?*

Answer: First, we have

$$f_Y(y) = f_X(x)\cdot J(x \to y)$$

where $f_X(x)$ is our $g(y_1)$. using this, we get

$$\begin{aligned} f_{Y_1(y_1)} &= \int_{y_2} g(y_1) J(x \to y_1, y_2)\, \mathrm{d}y_2 \\ &= g(y_1) \underbrace{\int_{y_2} J(x \to y_1, y_2)\, \mathrm{d}y_2}_{\text{Integral Jacobian}} \end{aligned}$$

We can also use the notation $J(x \to y_1)$ for the Integral Jacobian since we see that the dependence on y_2 will disappear as we are integrating over it. Hence,

$$f_{Y_1}(y_1) = g(y_1) \cdot J(x \to y_1)$$

There are two ways to use this

- To derive density of $f_{y_1}(y_1)$ provided $J(x \to y_1)$ is easy to compute
- To find $J(x \to y_1)$, provided $f_{Y_1}(y_1)$ and $g(y_1)$ are easy to find.

(Hsu's Lemma: $X \to S \equiv X'X$)

Our plan:

- First, derive Hsu's Lemma, where we derive $J(x \to s)$.
- Once we know $J(x \to s)$, we use the Jacobian to derive the density of the Wishart distribution

Example 9. *In polar coordinates,*

$$J(x \to \rho, \theta_1, \ldots, \theta_{n-1}) = \rho^{n-1} \prod_{i=1}^{n-1} (\sin\theta_i)^{n-i-1}$$

Therefore,

$$\begin{aligned} J(x \to \rho) &= \int_{\theta_1} \int_{\theta} \cdots \int_{\theta_{n-1}} \rho^{n-1} \prod_{i=1}^{n-1} (\sin\theta_i)^{n-i-1}\, \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_{n-1} \\ &= \rho^{n-1} \frac{2\pi^{n/2}}{\Gamma(n/2)} \end{aligned}$$

When $X \sim \mathcal{N}_p(\mathbf{0}, I_p)$,

$$f_X(x) = \frac{1}{(2\pi)^{\rho/2}} e^{-\rho^2/2} = g(y_1)$$

So

$$\begin{aligned} f_\rho(\rho) &= g(y_1) J(x \to y_1) \\ &\equiv \frac{1}{(2\pi)^{\rho/2}} e^{-\rho^2/2} \rho^{n-1} \frac{2\pi^{n/2}}{\Gamma(n/2)} \\ &= \frac{\rho^{\rho-1} e^{-\rho^2/2}}{2^{\frac{\rho-2}{2}} \Gamma(\rho/2)} \end{aligned}$$

Which is the density of $\sqrt{\chi^2_\rho(0)}$*, so*

$$\rho^2 \sim \chi^2_\rho(0)$$

B3.2 Jacobian to triangular coordinates. Let $V_{n\times p} = \underset{\sim}{X}' = (v_1, v_2, \ldots, v_p)$. As in A2.4 (Gram-Schmidt)

$$V_{n\times p} = W_{n\times p} T_{p\times p}$$

Then

$$X_{p\times p} \equiv \underset{\sim}{X}\underset{\sim}{X}' \equiv V'V = T'T$$

Now, the key is to derive integral Jacobian

$$X_{p\times n} \to T_{p\times p}$$

Idea: To compute $J(\underset{\sim}{X} \to T)$, could we come up with a special case where $f_{\underset{\sim}{X}}(\underset{\sim}{x})$ and $f_T(t)$ are both easy to compute, and also that

$$f_{\underset{\sim}{X}}(\underset{\sim}{x}) \equiv g(t)$$

for some g? Then

$$J(\underset{\sim}{x} \to t) = \frac{f_T(t)}{g(t)}$$

Let $X \sim \mathcal{N}_{p\times n}(\mathbf{0}, I_p \bigotimes I_n)$ where $X_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$.

- $$\begin{aligned} f_{\underset{\sim}{X}}(\underset{\sim}{x}) &= \frac{1}{(2\pi)^{np/2}} e^{-\frac{1}{2}\sum_{i=1}^{n} x_i' x_i} \\ &= \frac{1}{(2\pi)^{np/2}} e^{-\frac{1}{2}\mathrm{tr}(\underset{\sim}{X}\underset{\sim}{X}')} \\ &= \frac{1}{(2\pi)^{np/2}} e^{-\frac{1}{2}\mathrm{tr}(S)} \\ &= \frac{1}{(2\pi)^{np/2}} e^{-\frac{1}{2}\mathrm{tr}(T'T)} \\ &= \frac{1}{(2\pi)^{np/2}} e^{-\frac{1}{2}\sum_{i\le j} t_{ij}^2} \\ &\equiv g(t) \end{aligned}$$

- How do we calculate $f_T(t)$? We claim that all t_{ij} are independent and
$$t_{jj}^2 \sim \chi_{n-j+1}^2 \quad t_{ij} \sim \mathcal{N}(0,1)$$
for $i < j$.
This gives
$$\begin{aligned} f_T(t) &= \prod_{i=1}^{p} \frac{t_{ii}^{n-i} e^{-t_{ii}^2/2}}{2^{\frac{n-i-1}{2}} \Gamma\left(\frac{n+1-i}{2}\right)} \prod_{i=1}^{p-1} \prod_{j=i+1}^{p} \left(\frac{e^{-t_{ij}^2/2}}{\sqrt{2\pi}}\right) \\ &= c \left(\prod_{i=1}^{p} t_{ii}^{n-i}\right) e^{-\frac{1}{2}\sum_{i\le j} t_{ij}^2} \end{aligned}$$
Now we have
$$\begin{aligned} J(\underset{\sim}{X} \to T) &= \frac{f_T(t)}{g(t)} \\ &= c_1 \prod_{i=1}^{p} t_{ii}^{n-i} \end{aligned}$$
where
$$c_1 = \frac{2^p \pi^{\frac{np}{2} - \frac{p^2}{4} + \frac{p}{4}}}{\prod_{j=1}^{p} \Gamma\left(\frac{n-j+1}{2}\right)}$$

Lemma (Lemma 1)**.** *The integral Jacobian from $\underset{\sim}{X}_{p\times n} \to T$ is*
$$J(\underset{\sim}{x} \to T) = c_1 \prod_{i=1}^{p} t_{ii}^{n-i}$$
where c_1 is defined as above

Proof.

(1) $V = WT$, so $V_1 = W_1 \cdot t_{11}$, where
$$t_{11}^2 = \|V_1\|^2 \sim \chi_n^2(0)$$

(2)
$$\begin{aligned} V_2 &= t_{12} W_1 + t_{22} W_2 \\ &= \hat{V}_2 + \overset{\perp}{V}_2 \end{aligned}$$
with $t_{12} = W_1' V_2$ and $t_{22}^2 = \|\overset{\perp}{V}_2\|^2$
$$\begin{aligned} P &= P_{W_1} \\ &= W_1 (W_1' W_1)^{-1} W_1' \\ &= W_1 W_1' \end{aligned}$$
$$\|\overset{\perp}{V}_2\|^2 = \|(I-P)V_2\|^2 \sim \chi_{n-1}^2(0)$$

(a) when $P = P_{W_1}$, with W_1 nonstochastic,
$$\|(I-P)V_2\|^2 \sim \chi_{n-1}^2(0)$$

(b) Show since V_1, V_2 are independent

$$\|(I-P)V_2\|^2 \sim \chi^2_{n-1}(0)$$

$$\int f(t|v_1)f(v_1)\,\mathrm{d}v_1 = \int f(y)f(v_1)\,\mathrm{d}v_1$$

(3)

$$t_{12} = W_1'V_2$$
$$= \begin{cases} \text{if } W_1 \text{ is nonstochastic, } \|W_1\| = 1, \ \sim \mathcal{N}(0,1) \\ \text{now, } W_1 \text{ and } V_2 \text{ are independent so } \ \sim \mathcal{N}(0,1) \end{cases}$$

Independence:

$$t_{12} = W_1'PV_2$$
$$t_{22}^2 = \|\overset{\perp}{P}V_2\|^2$$

□

Homework (B3.2)**.**

(a) *Carry out step 3 meaning*

$$V_3 = t_{13}W_1 + t_{23}W_2 + t_{33}W_3$$

Show that t_{13}, t_{23}, t_{33}*, which are respectively* $\mathcal{N}(0,1), \mathcal{N}(0,1), \chi^2_{n-2}$*, are independent of* t_{11}, t_{12}, t_{22}*.*

(b) *Explicitly verify* c_1

Corollary. *If*

$$\underset{\sim}{X} \sim \mathcal{N}_{p\times n}(\mathbf{0}, \Sigma \bigotimes I_n)$$

Then

$$f_T(t) = \frac{c_1(2\pi)^{-np/2}}{|\Sigma|^{n/2}} \prod_{i=1}^{p} t_{ii}^{n-i} e^{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}S)}$$

where $S = T'T$

Proof.

$$f_T(t) = f_{\underset{\sim}{X}}(\underset{\sim}{x}) J(\underset{\sim}{X} \to T)$$

where $f_X(x)$ is $g(t)$ for some t

$$\left(\frac{(2\pi)^{-np/2}}{|\Sigma|^{n/2}} e^{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}S)}\right) g(t)$$

□

B3.3. We have Mapping $T \to S$
We hope: Jacobian $\underset{\sim}{X} \to S$
So far: Jacobian $\underset{\sim}{X} \to T$
What remains: Jacobian $T \to S$

$$J(\underset{\sim}{X} \to S) = J(\underset{\sim}{X} \to T) J(T \to S)$$

Lemma (Lemma 2)**.** *If* $x \to (y_1, y_2)$ *and* $y_1 \to (z_1, z_2)$*, then the chain rule holds,*

$$J(x \to z_1) = J(x \to y_1) J(y_1 \to z_1)$$

Proof. The mapping $T \to S \equiv T'T$ takes the upper triangular matrix $T_{p\times p}$ to a $p \times p$ symmetric matrix □

Lemma (Lemma 3)**.**

$$J(T \to S) = \left(2^p \prod_{j=1}^{p} t_{jj}^{p-j+1}\right)^{-1}$$

Homework (B3.3)**.** *Verify Lemma 3.*

Hint:

- Use induction in p
- Examine $\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}$ matrix

$$\left(\frac{\partial s_{ij}}{\partial t_{kl}}\right) = \begin{bmatrix} A & 0 \\ B & C \end{bmatrix}$$

9. October 2nd, 2013

Last time we had that the Integral Jacobian had the formula

$$f_T(t) = g(t)J(\underset{\sim}{X} \to T) \equiv F_{\underset{\sim}{X}}(x)$$

- "$\Rightarrow$" We use this to find $J(\underset{\sim}{X} \to T)$
- "$\Leftarrow$" We use $J(\underset{\sim}{X} \to T)$ to find $f_T(t)$

Two main results:

(a) $J(\underset{\sim}{X} \to T) = c_1 \prod_{i=1}^{p} t_{ii}^{n-i}$ where

$$V \equiv X' = WT$$

and

$$c_1 = \frac{2^p \pi^{\left(\frac{np}{2} - \frac{p^2}{4} + \frac{p}{4}\right)}}{\prod_{j=1}^{p} \Gamma\left(\frac{n-j+1}{2}\right)}$$

(b) $J(T \to S) = \left[2^p \prod_{j=1}^{p} t_{jj}^{p-j+1}\right]^{-1}$

Lemma (Hsu's Lemma)**.**

$$J(\underset{\sim}{X} \to S) = c_2 |S|^{\frac{1}{2}(n-p-1)}$$

where

$$c_2 = \frac{\pi^{\frac{np}{2} - \frac{p^2}{4} + \frac{p}{4}}}{\prod_{j=1}^{p} \Gamma(\frac{n-j+1}{2})}$$

Proof.

$$J(\underset{\sim}{X} \to S) = J(\underset{\sim}{X} \to T)J(T \to S)$$

□

Homework (B3.4)**.** *check this*

9.0.1. *Geometric Interpretation.*

$$|S| = |X'X| = |V'V|$$

where $v = (v_1 \dots v_p)$ and

$$|S| = \left[\text{Vol}(v_1 \dots v_p)\right]^2$$

By Hsu's Lemma,

$$J(\underset{\sim}{X} \to S) = c_2 \left[\text{Vol}(v_1 \dots v_p)\right]^{n-p-1}$$

where $n \geq p + 1$.

Homework (B3.5)**.** *Vector $v_1, \dots, v_p$ are chosen independently and uniformly on the unit sphere in $\mathbb{R}^n$. For positive integer a, find*

$$\mathbb{E}\left[\text{Vol}(v_1, \dots, v_p)^{2a}\right]$$

If $v_1^*, \dots, v_p^* \sim \mathcal{N}_n(\mathbf{0}, I_n)$, then

$$v_i = \frac{v_i^*}{\|v_i^*\|}, \qquad i = 1, \dots, p$$

Then

$$\begin{aligned}\left[\text{Vol}(v_1, \dots, v_p) \dot{:} (\|v_1^*\| \dots \|v_p^*\|)\right]^2 &= \left[\text{Vol}(v_1^* \dots v_p^2)\right]^2 \\ &= |S| \\ &= \prod_{i=1}^{p} t_{ii}^2\end{aligned}$$

Also

$$\begin{aligned}&\mathbb{E}\Big[\left[\text{Vol}(v_1 \dots v_p)\right]^{2a}\Big] \\ &= \mathbb{E}\left[\left[\|v_1^*\| \dots \|v_p^*\|\right]^{2a}\right] \\ &= \mathbb{E}\left[\text{Vol}(v_1^*, \dots, v_p^*)^{2a}\right] \\ &= \prod_{i=1}^{p} t_{ii}^{2a}\end{aligned}$$

where $t_{ii}^2 \sim \chi_{n-i+1}^2$

B4.1 Wishart.

Definition 9 (Wishart)**.** *If*

$$\underset{\sim}{X} \sim \mathcal{N}_{p\times n}(\mathbf{0}, \Sigma_{p\times p} \bigotimes I_n)$$

then

$$S = \underset{\sim}{X}\underset{\sim}{X}' = \sum_{i=1}^{n} x_i x_i'$$

(where $\underset{\sim}{X} = [x_1 \dots x_n]$ is said to have Wishart distribution,

$$S \sim W(\Sigma; n, p)$$

Theorem 4. *If $n \geq p$ and $\Sigma > 0$, then $S > 0$ with probability 1. For $\Sigma = I_p$,*

$$|S| = \prod_{i=1}^{p} t_{ii}^2$$

We also have that the density (which is a generalization of χ^2)

$$f_S(s) = c_3 |s|^{\frac{n-p-1}{2}} e^{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}S)} |\Sigma|^{-n/2}$$

for $s \geq 0$, where

$$c_3 = \left[2^{\frac{np}{2}} \pi^{\frac{p}{4}(p-1)} \prod_{j=1}^{p} \Gamma\left(\frac{n-j+1}{2}\right) \right]^{-1}$$

Proof.

$$f_{\underset{\sim}{X}}(\underset{\sim}{x}) = (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \exp\Bigg(-\frac{1}{2} \underbrace{\sum_{i=1}^{n} x_i' \Sigma^{-1} x_i}_{\mathrm{tr}(\Sigma^{-1}s)} \Bigg)$$

Since

$$\begin{aligned} \sum_{i=1}^{n} x_i' \Sigma^{-1} x_i &= \mathrm{tr}\left(\sum_{i=1}^{n} \Sigma^{-1} x_i x_i' \right) \\ &= \mathrm{tr}\left(\Sigma^{-1} \sum_{i=1}^{n} x_i x_i' \right) \\ &= \mathrm{tr}(\Sigma^{-1} s) \end{aligned}$$

This implies

$$\begin{aligned} f_S(s) &= \underbrace{f_{\underset{\sim}{X}}(\underset{\sim}{x})}_{=g(s)} J(\underset{\sim}{X} \to S) \\ &= (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} e^{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}s)} c_2 |s|^{\frac{n-p-1}{2}} \\ &\equiv c_3 |s|^{\frac{n-p-1}{2}} |\Sigma|^{-\frac{n}{2}} e^{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}s)} \end{aligned}$$

□

Properties and Facts:

(1) If $S_{p\times p} \sim W(\Sigma; n, p)$ and $\tilde{S} = ASA'$ where $\tilde{S}$ is $q \times q$, A is $q \times p$, S is $p \times p$, and A' is $p \times q$, then

$$\tilde{S} = W(A\Sigma A'; n, q)$$

(2) If $S_1 \sim W(\Sigma; n_1, p) \perp\!\!\!\perp S_2 \sim W(\Sigma; n_2, p)$, then

$$S_1 + S_2 \sim W(\Sigma; n_1 + n_2, p)$$

(3) If $\underset{\sim}{X}_{p\times n} \sim \mathcal{N}_{p\times n}(U, \Sigma \bigotimes I_n)$ and $\Gamma_{n\times m}$ such that

$$\underbrace{U}_{p\times n}\underbrace{\Gamma}_{n\times m} = \underbrace{\mathbf{0}}_{p\times n}$$

then, if we let

$$\underbrace{\underset{\sim}{Y}}_{p\times m} = \underbrace{\underset{\sim}{X}}_{p\times n}\underbrace{\Gamma}_{n\times m}$$

then

$$\underset{\sim}{Y}\underset{\sim}{Y}' \sim W(\Sigma; m, p)$$

Proof.

$$\underset{\sim}{Y} = \underset{\sim}{X}\Gamma \sim \mathcal{N}(\underbrace{U\Gamma}_{0}, \Sigma \bigotimes \underbrace{\Gamma\Gamma'}_{I_m})$$

□

(4) If

$$\underbrace{\underset{\sim}{Y}}_{p\times n} = \underbrace{\underset{\sim}{X}}_{p\times n}\underbrace{P}_{n\times n}$$

where $P = \Gamma\Gamma'$, which is $n \times n$ is the projection matrix to $L_{col}(P)$ and $UP = 0$. Then

$$\underset{\sim}{Y}\underset{\sim}{Y}' = \underset{\sim}{X}P^2\underset{\sim}{X}' = \underset{\sim}{X}P\underset{\sim}{X}' = (\underset{\sim}{X}\Gamma)(\Gamma'\underset{\sim}{X}')$$

which implies

$$\underset{\sim}{Y}\underset{\sim}{Y}' \sim W(\Sigma; m, p)$$

(5) Decomposition Property:

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \overset{iid}{\sim} \mathcal{N}_{p+q}\left(\mathbf{0}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}\right)$$

where Σ_{XX} is $p \times p$ and Σ_{YY} is $q \times q$ and $\Sigma_{XX} > 0$. Then the Data matrix

$$\begin{bmatrix} \underset{\sim}{X} \\ \underset{\sim}{Y} \end{bmatrix} \sim \mathcal{N}_{(p+q)\times n}(\mathbf{0}, \Sigma \bigotimes I_n)$$

where $\underset{\sim}{X}$ is $p \times n$ and $\underset{\sim}{Y}$ is $q \times n$ and also

$$W = \begin{bmatrix} \underset{\sim}{X} \\ \underset{\sim}{Y} \end{bmatrix} \begin{bmatrix} \underset{\sim}{X}' & \underset{\sim}{Y}' \end{bmatrix} = \begin{bmatrix} S_{XX} & S_{XY} \\ S_{YX} & S_{YY} \end{bmatrix} \sim W(\Sigma; n, p+q)$$

Homework (B4.1)**.** *Suppose A is non-singular. Show that the mapping $S \to ASA'$ has Jacobian*

$$J(S \to \tilde{S}) = \left(|A|^2\right)^{-\frac{p+1}{2}}$$

Hint: $f_{\tilde{S}}(\tilde{s})$ and $f_S(s)$

Next Problem:
In Wishart, $X_{p,n} \sim \mathcal{N}_{p,n}(0, \Sigma \bigotimes I_n)$. Let $P = I_n - \frac{\mathbf{1}_n\mathbf{1}_n'}{n}$. If $X_2 \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$, then $\underset{\sim}{X} \sim \mathcal{N}_{p\times n}(u\mathbf{1}_n', \Sigma \bigotimes I_n)$.

$$\mu\mathbf{1}_n' \cdot P = \mu \cdot (P \cdot \mathbf{1}_n') = \underbrace{\mathbf{0}}_{p\times n}$$

So

$$\begin{aligned} \underset{\sim}{Y} &= \underset{\sim}{X}P \\ &\equiv [x_1, \dots x_n]P \\ &= (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}) \end{aligned}$$

which implies

$$\underset{\sim}{Y}\underset{\sim}{Y}' \sim W(\Sigma; n-1, p) \Leftrightarrow \sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'$$

Homework (B4.2)**.** *Let*

$$\underset{\sim}{X} = \underbrace{Q}_{p\times p}\underbrace{X}_{p\times n}\underbrace{P}_{n\times n}$$

where

$$Q = I_p - \frac{\mathbf{1}_p\mathbf{1}_p'}{p}$$

$$P = I_n - \frac{\mathbf{1}_n\mathbf{1}_n'}{n}$$

Show that $\tilde{S}$ (i.e. $\underset{\sim}{X}\underset{\sim}{X}'$) $\sim W(\tilde{\Sigma}; n-1, p)$ where the elements of $\tilde{\Sigma}$ are

$$\tilde{\sigma}_{ij} = \sigma_{ij} - \sigma_{i*} - \sigma_{*j} + \sigma_{**}$$

Then $\tilde{Y}\tilde{Y}' = \underset{\sim}{X}P^2\underset{\sim}{X}' = \underset{\sim}{X}\underset{\sim}{X}' = S_{XX}$ implies

$$\tilde{Y}\tilde{Y}' \sim W(\Sigma; m, p)$$

Define $\hat{P}_X = \underset{\sim}{X}'(\underset{\sim}{X}\underset{\sim}{X}')^{-1}\underset{\sim}{X}$ which is $n \times n$. This is the projection matrix to $L_{row}(\underset{\sim}{X})$ and

$$\begin{aligned}\overset{\perp}{P}_X &= I_n - \hat{P}_X \\ \underset{\sim}{Y} &= \underbrace{\hat{Y}_X}_{\underset{\sim}{Y}\hat{P}_X} + \underbrace{\overset{\perp}{Y}_X}_{\underset{\sim}{Y}\overset{\perp}{P}_X}\end{aligned}$$

where $\underset{\sim}{Y}$ is $q \times n$. Likewise, decompose

$$\begin{aligned}S_{YY} &= \underset{\sim}{Y}\underset{\sim}{Y}' \\ &= (\hat{Y}_X + \overset{\perp}{Y}_X)(\hat{Y}_X + \overset{\perp}{Y}_X)' \\ &= \hat{Y}_X\hat{Y}_X' + \overset{\perp}{Y}_X\overset{\perp}{Y}_X'\end{aligned}$$

For the first term, we have

$$\begin{aligned}\underset{\sim}{Y}_X\hat{P}_X\underset{\sim}{Y}_X' &= \underset{\sim}{Y}\underset{\sim}{X}'(\underset{\sim}{X}\underset{\sim}{X}')^{-1}\underset{\sim}{X}\underset{\sim}{Y}' \\ &= S_{YX}S_{XX}^{-1}S_{XY}\end{aligned}$$

Now,

$$\begin{aligned}\overset{\perp}{S}_{YY} &= \overset{\perp}{Y}_X\overset{\perp}{Y}_X' \\ &= S_{YY} - S_{YX}S_{XX}^{-1}S_{XY}\end{aligned}$$

(a) $S_{XX} \sim W(\Sigma_{XX}; n, p)$ (trivial)

(b) $(S_{YX}|S_{XX}) \sim \mathcal{N}_{q\times p}(\Sigma_{YX}\Sigma_{XX}^{-1}S_{XX}, \overset{\perp}{\Sigma}_{YY} \bigotimes S_{XX})$ where

$$\overset{\perp}{\Sigma}_{YY} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$

(c) $(\overset{\perp}{S}_{YY}|S_{XX}) \sim W(\overset{\perp}{\Sigma}_{YY}; n-p, q)$

(d) S_{YX} and $\overset{\perp}{S}_{YY}$, the two parts of S_{YY} are conditionally independent given S_{XX}. Also $\overset{\perp}{S}_{YY} \perp\!\!\!\perp S_{XX}$.
The proof is since $(\overset{\perp}{S}_{YY}|S_{XX})$ does not depend on S_{XX}

10. October 7th, 2013

Useful trick??

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{a}{z_0}+z_0\right)^2}\,\mathrm{d}z_0$$

which is

$$C\int_0^{\infty} e^{-\frac{1}{2}\left(\frac{1}{z_0}+z_0\right)^2}\,\mathrm{d}z_0 + a\int_0^{\infty} e^{-\frac{1}{2}\left(\frac{1}{z_0}+z_0\right)^2}\,\mathrm{d}\left(\frac{1}{z_0}\right) = \int_0^{\infty} e^{-\frac{1}{2}\left(\frac{1}{z_0+z_0}\right)^2}\,\mathrm{d}\left(z_0+\frac{1}{z_0}\right)$$

Decomposition Lemma. From last time,

- $$\begin{pmatrix}X_i \\ Y_i\end{pmatrix} \overset{iid}{\sim} \mathcal{N}_{p+q}\left(\mathbf{0}, \begin{bmatrix}\Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY}\end{bmatrix}\right)$$
  where Σ_{XX} is $p \times p$ and Σ_{YY} is $q \times q$ and $\Sigma_{XX} > 0$.
- Then the Data matrix is
  $$\begin{bmatrix}\underset{\sim}{X} \\ \underset{\sim}{Y}\end{bmatrix} \sim \mathcal{N}_{(p+q)\times n}(\mathbf{0}, \Sigma \bigotimes I_n)$$
  where $\underset{\sim}{X}$ is $p \times n$ and $\underset{\sim}{Y}$ is $q \times n$.
- The Wishart matrix is
  $$S = \begin{bmatrix}S_{XX} & S_{XY} \\ S_{YX} & S_{YY}\end{bmatrix} \equiv \begin{bmatrix}\underset{\sim}{X}\underset{\sim}{X}' & \underset{\sim}{X}\underset{\sim}{Y}' \\ \underset{\sim}{Y}\underset{\sim}{X}' & \underset{\sim}{Y}\underset{\sim}{Y}'\end{bmatrix}$$
- Random projection:

$$\begin{aligned}\hat{P}_X &= \underset{\sim}{X}'(\underset{\sim}{X}\underset{\sim}{X}')^{-1}\underset{\sim}{X} \\ \overset{\perp}{P}_X &= I_n - \hat{P}_X \\ \underset{\sim}{Y} &= \underset{\sim X}{\hat{Y}} + \underset{\sim X}{\overset{\perp}{Y}} \\ S_{YY} &= S_{YX}S_{XX}^{-1}S_{XY} + \overset{\perp}{S}_{YY}\end{aligned}$$

where

$$S_{YX}S_{XX}^{-1}S_{XY} = \hat{\underset{\sim}{Y}}_X \underset{\sim}{Y}'_X$$
$$\overset{\perp}{S}_{YY} = \underset{\sim}{Y}_X \underset{\sim}{Y}'_X$$

??

Lemma (Decomposition Lemma)**.**

(1) $S_{XX} \sim W(\Sigma_{XX}; n, p)$ *(trivial)*

(2) $S_{XY}|S_{XX} \sim \mathcal{N}_{q\times p}(\Sigma_{YX}\Sigma_{XX}^{-1}S_{XX}, \overset{\perp}{\Sigma}_{YY} \bigotimes S_{XX})$

Proof. $Y_i|X_i \overset{iid}{\sim} \mathcal{N}_q(\Sigma_{YX}\Sigma_{XX}^{-1}X_i, \overset{\perp}{\Sigma}_{YY})$ therefore,

$$\underset{\sim}{Y}_{q\times n}|\underset{\sim}{X}_{p\times n} \sim \mathcal{N}_{q\times n}(\Sigma_{YX}\Sigma_{XX}^{-1}\underset{\sim}{X}, \overset{\perp}{\Sigma}_{YY} \bigotimes I_n)$$

This says given $\underset{\sim}{X}$,

$$\hat{\underset{\sim}{Y}}_X \equiv \underset{\sim}{Y}\hat{P}_X \sim \mathcal{N}_{q\times n}(\Sigma_{YX}\Sigma_{XX}^{-1}\underset{\sim}{X}\hat{P}_X, \overset{\perp}{\Sigma}_{YY} \bigotimes \hat{P}_X^2)$$

since

$$S_{YX} \equiv \underset{\sim}{Y}\underset{\sim}{X}' \equiv (\hat{Y}_X + \overset{\perp}{Y}_X)\underset{\sim}{X}' \equiv \hat{Y}_X\underset{\sim}{X}'$$
$$S_{YX}|\underset{\sim}{X} \sim \mathcal{N}_{q\times p}(\Sigma_{YX}\Sigma_{XX}^{-1}\underbrace{\underset{\sim}{X}\hat{P}_X\underset{\sim}{X}'}_{S_{XX}}, \overset{\perp}{\Sigma}_{YY} \bigotimes \underbrace{\underset{\sim}{X}\hat{P}_X^2\underset{\sim}{X}'}_{S_{XX}}) \qquad (\star)$$

Note that $S_{YX}|\underset{\sim}{X}$ has the same distribution as $S_{YX}|S_{XX}$ because the right hand side of $\star$ only depends on S_{XX}. □

(3) $\overset{\perp}{S}_{YY}|S_{XX} \sim W(\overset{\perp}{\Sigma}_{YY}, n-p, q)$

Proof. Similarly,

$$\overset{\perp}{Y}_X|\underset{\sim}{X} \sim \mathcal{N}_{q\times n}(\underbrace{\Sigma_{YX}\Sigma_{XX}^{-1}\underset{\sim}{X}\overset{\perp}{P}_X}_{=0}, \overset{\perp}{\Sigma}_{YY} \bigotimes \overset{\perp}{P}_X^2)$$

We have

$$\overset{\perp}{Y}_X = Y\overset{\perp}{P}_X$$

where $Y \sim \mathcal{N}(\mu, \Sigma \bigotimes I_n)$, so $\mu\overset{\perp}{P}_X = 0$. Use (iv) of B4.1, Page 6, which gives us

$$\overset{\perp}{S}_{YY} = \overset{\perp}{\underset{\sim}{Y}}\overset{\perp}{\underset{\sim}{Y}}' \sim W(\overset{\perp}{\Sigma}_{YY}, n-p, q)$$

□

(4) $S_{YX}S_{XX}^{-1}S_{XY}$ *and* $\overset{\perp}{S}_{YY}$*, the two parts of* S_{YY} *are conditionally independent, given* S_{XX}*. (To prove this, all you need to show is that* $S_{XY}|S_{XX}$ *and* $\overset{\perp}{S}_{YY}|S_{XX}$ *are independent.)*

B5.1 Functions of Wishart. If $S \sim W(\Sigma; n, p)$ with $\Sigma > 0$, then what is $\tilde{S} \equiv S^{-1}$? We call this the Inverse of the Wishart ($IW(\Sigma; n, p)$).

Lemma. *The mapping* $S \to S^{-1}$ *of the symmetric positive definite matrices to itself has Jacobian*

$$J(S \to \underbrace{\tilde{S}}_{=S^{-1}}) = |S|^{p+1}$$

Recall from last time that if A is fixed, nonsingular and $p \times p$, then

$$J(S \to ASA') = |A|^{p+1}$$

"Proof". If we add a small perturbation to S, $S \to S + \mathrm{d}S$, where $\mathrm{d}S$ is a small perturbation symmetric matrix, then S^{-1} is also going to be perturbed

$$S^{-1} = T \to T + \mathrm{d}T$$

such that

$$(S + \mathrm{d}S)(T + \mathrm{d}T) = I_p$$

Now consider

$$(S + \mathrm{d}S)(S^{-1} - S^{-1}(\mathrm{d}S)S^{-1}) = I_p + \underbrace{\mathrm{d}SS^{-1} - \mathrm{d}SS^{-1}}_{=0} - \underbrace{(\mathrm{d}S)S^{-1}(\mathrm{d}S)S^{-1}}_{\text{higher order terms}}$$

Therefore,

$$\mathrm{d}T \approx -S^{-1} \cdot \mathrm{d}S \cdot S^{-1} = -T \cdot \mathrm{d}S \cdot T$$

which acts as a mapping

$$\mathrm{d}S \to \mathrm{d}T \equiv -T \cdot \mathrm{d}S \cdot T$$

□

Using this, we can think of the Jacobian $J(S \to \tilde{S}) = |T|^{-(p+1)} = |S|^{p+1}$.

Corollary. *If $S \sim W(\Sigma; n, p)$ with $\Sigma > 0$, then $T = S^{-1}$ has density*

$$f_T(t) = c_3 |t|^{-\frac{n+p+1}{2}} \frac{e^{-\frac{1}{2}\mathrm{tr}\Sigma^{-1}t^{-1}}}{|\Sigma|^{n/2}}$$

with

$$c_3 = \left[2^{\frac{np}{2}} \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \Gamma\left(\frac{n-j+1}{2} \right) \right]^{-1}$$

Homework (B5.1)**.** *Use the Lemma and Wishart density to verify the Corollary.*

Correlation Matrix. Let $D = \mathrm{diag}(d_1, \ldots, d_p)$, where $d_i = \sqrt{s_{ii}}$

$$R = D^{-1}SD = \begin{cases} 1 & \text{diagonals} \\ \gamma_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} & \text{off-diagonal} \end{cases}$$

$S = \underset{\sim}{X}\underset{\sim}{X}'$ and $S_{ij} = v_i' v_j = \sum_{k=1}^{n} X_{ik}^2$, which is called the empirical correlation. The question to ask is what is the distribution of (R, D)?

Lemma. $J(S \to R, D) = 2^p |D|^p$.

Homework (5.2)**.** *Verify this. (All it takes it a normalization by a diagonal matrix).*

Corollary. *If $S \sim W(I_p, n, p)$, then D and R are independent, with*

$$f_R(r) = c_4 |R|^{\frac{n-p-1}{2}}$$

and

$$d_i^2 \overset{iid}{\sim} \chi_n^2(0)$$

Proof.

$$\begin{aligned} f(D, R) &= f_S(s) J(S \to D, R) \\ &= c_3 |D|^p |S|^{\frac{n-p-1}{2}} e^{-\frac{1}{2}\mathrm{tr}S} \quad \text{because } \Sigma = I_p \end{aligned}$$

Hence $|D|^p$ is the Jacobian since the rest is the Wishart. Furthermore, $|S| = |D|^2|R|$. So $|S|$ decomposes into a product of something in terms of D and something in terms of R. What do we do with the trace?

$$\begin{aligned} \mathrm{tr}(S) &= \mathrm{tr}(DRD) \\ &= \mathrm{tr}(D^2 R) \end{aligned}$$

where D^2 is a diagonal matrix and R is a matrix with unit diagonals. Hence

$$\mathrm{tr}(S) = \mathrm{tr}(D^2)$$

So

$$f(D, R) = c_3 |D|^{n-1} |R|^{\frac{n-p-1}{2}} e^{-\frac{1}{2}\mathrm{tr}(D^2)}$$

This doesn't prove everything though □

Homework (B5.3)**.** *Finish the proof and "what goes wrong" if $\Sigma \neq I_p$?*

B5.2 Hotelling's T^2.

One-sample t (Univariate Case). Given $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where σ^2 is unknown (μ is unknown). Our hypotheses are $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$.
Geometry: $\mathbb{R}^n$. Consider

$$\underset{\sim}{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \mu \cdot \mathbf{1}_n + \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}$$

where $Z_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. We project $\underset{\sim}{x}$, which gives us $\hat{\underset{\sim}{x}}$ and $\overset{\perp}{\underset{\sim}{x}}$. The angle $\hat{A}$ is the angle between $\underset{\sim}{x}$ and $\hat{\underset{\sim}{x}}$. If $\hat{A}$ is small, we reject. If it is large, we accept.

In other words, we reject when

$$\mathrm{cotan}^2(\hat{A}) = \frac{\|\hat{X}\|^2}{\|\overset{\perp}{X}\|^2}$$

is large. This is equivalent to when

$$t^2 = \frac{\|\hat{X}\|^2}{\|\overset{\perp}{\hat{X}}\|^2/(n-1)} = \frac{n\bar{x}^2}{\sum(x_i-\bar{x})^2/(n-1)} \equiv (n-1)\text{cotan}^2(\hat{A})$$

In p-dimensions, with $X_i \overset{iid}{\sim} \mathcal{N}_p(\mu, \Sigma)$, we want to test $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$.

Consider $L_{row}(X)$. When $p = 1$, we only have one $n \times 1$ vector. When $p > 1$, we have p $n \times 1$ vectors. The space spanned by $(v_1, \ldots, v_p)$ is "uniform" over all p-dimensional subspaces of $\mathbb{R}^n$ when $\mu = 0$.

Project $\mathbf{1}_n$ to $L_{row}(\underset{\sim}{X})$.

$$\underset{\sim}{X} = \begin{pmatrix} \vdots & \vdots & \vdots \\ X_1 & \cdots & X_n \\ \vdots & \vdots & \vdots \end{pmatrix} = \begin{pmatrix} \cdots & v_1 & \cdots \\ \ldots & \vdots & \ldots \\ \cdots & v_p & \cdots \end{pmatrix}$$

If $\mu = 0$, the space of "$v_1 \ldots v_p$" is random and we expect a large angle.

$$\mathbf{1}_n = a + b$$

where $a \in L_{row}(X)$ and $b \in \overset{\perp}{L}_{row}(X)$. Therefore,

$$T^2 \propto \frac{\|a\|^2}{\|b\|^2}$$

11. October 9th, 2013

B6.1 One-sample t-statistic. (Univariate) Observe $X_i, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ and we wish to test

$$H_0 : \mu = 0 \qquad H_1 : \mu \neq 0$$

One sided:

$$t = \frac{\sqrt{n}\bar{X}}{\sqrt{S/(n-1)}}$$

where $S = \sum_{i=1}^n (X_0 - \bar{X})^2$. We reject when $t \geq t_0$, $t \sim t_{n-1}(0)$, $t_0 = t_{n-1}^{\alpha}$.

In the two-sided t-test,

$$\begin{aligned} t^2 &= \frac{n\bar{X}^2}{S/(n-1)} \\ &= \frac{\|\hat{X}\|^2}{\|\overset{\perp}{\hat{X}}\|^2/(n-1)} \\ &= (n-1)\text{cotan}^2(\hat{A}) \end{aligned}$$

where the last two lines are the geometric interpretation.

Rationale:

- $\frac{x}{\|x\|}$ is uniform over the $(n-1)$-dimensional sphere in $\mathbb{R}^n$.
- $\frac{\mathbf{1}_n}{\|\mathbf{1}_n\|}$ is fixed.

If $\mu \neq 0$, then

$$x = \mu \cdot \mathbf{1}_n + Z, \quad Z \sim \mathcal{N}(\mathbf{0}, I_n)$$

Hotelling developed the analogous theory for the multivariate case, testing

$$H_0 : \mu_{p\times 1} = 0 \text{ vs } H_1 : \mu_{p\times 1} \neq 0$$

Given the data $X_i \overset{iid}{\sim} \mathcal{N}_p(\mu, \Sigma)$.

$$\underset{\sim}{X} = \begin{pmatrix} \vdots & \vdots & \vdots \\ x_1 & \cdots & x_n \\ \vdots & \vdots & \vdots \end{pmatrix} = \begin{pmatrix} \cdots & v_1' & \cdots \\ \ldots & \vdots & \ldots \\ \cdots & v_p' & \cdots \end{pmatrix}$$

- $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \sim \mathcal{N}(\mu, \Sigma/n)$
-

$$\begin{aligned} S &= \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})' \\ &\equiv \underset{\sim}{X}\underset{\sim}{X}' - n\bar{\underset{\sim}{X}}\bar{\underset{\sim}{X}}' \\ &\sim W(\Sigma; n-1, p) \end{aligned}$$

- $\bar{X}$ and S are independent

$$P = \frac{\mathbf{1}_n \mathbf{1}_n'}{n} \qquad \bar{X} = \underset{\sim}{X} P \qquad S = (I_n - P)$$

Definition 10 (Hotelling's T^2-test)**.** *Hotelling's T^2-test is*

$$T^2 = \underbrace{\bar{X}'}_{1\times p} \underbrace{\left[\frac{S}{n(n-1)}\right]^{-1}}_{p\times p} \underbrace{\bar{X}}_{p\times 1} = \bar{X}' \left[\widehat{\mathrm{Var}}(\bar{X})\right]^{-1} \bar{X}$$

where

$$\bar{X} \sim \mathcal{N}\left(\mathbf{0}, \frac{\Sigma}{n}\right)$$

if H_0 holds.

$$\frac{S}{n(n-1)}$$

is an unbiased estimate of $\frac{\Sigma}{n}$.

Geometric Interpretation: Let $\hat{A}$ be the minimum angle between $L(\mathbf{1}_n)$ and any vector in $L_{row}(X)$.

Claim.

$$T^2 = (n-1)\mathrm{cotan}^2(\hat{A})$$

Proof. Applying (vii) on A2.3,

$$\begin{aligned}
\cos^2(\hat{A}) &= \frac{\|P_X \mathbf{1}_n\|^2}{\|\mathbf{1}_n\|^2} \\
&= \frac{1}{n}\mathbf{1}_n' P^2 \mathbf{1}_n \\
&= \frac{1}{n}\mathbf{1}_n' P \mathbf{1}_n \\
&= \frac{1}{n}\mathbf{1}_n' \underset{\sim}{X}' (\underset{\sim}{X}\underset{\sim}{X}')^{-1} \underset{\sim}{X} \mathbf{1}_n \\
&= n\bar{X}' (\underset{\sim}{X}\underset{\sim}{X}')^{-1} \bar{X} \\
&= n\bar{X}' \left[S + n\bar{X}\bar{X}'\right]^{-1} \bar{X}
\end{aligned}$$

Now, use Woodbury's theorem (A4.1, class 4)

$$\begin{aligned}
\cos^2(\hat{A}) &= n\bar{X}' \left[S^{-1}\left(S - \frac{\bar{X}\bar{X}'}{\frac{1}{n} + \bar{X}'S^{-1}\bar{X}}\right) S^{-1}\right] \bar{X} \\
&= n\bar{X}' S^{-1} S S^{-1} \bar{X} - \frac{n^2 (\bar{X}'S^{-1}\bar{X})(\bar{X}'S^{-1}\bar{X})}{n\left[\frac{1}{n} + \bar{X}'S^{-1}\bar{X}\right]} \\
&= \underbrace{n\bar{X}'S^{-1}\bar{X}}_{a^2} - \frac{\underbrace{(n\bar{X}'S^{-1}\bar{X})^2}_{a^4}}{\underbrace{1 + n\bar{X}'S^{-1}\bar{X}}_{1+a^2}} \\
&= \frac{a^2}{1+a^2}
\end{aligned}$$

which implies

$$\cos^2(\hat{A}) = \frac{a^2}{1+a^2}$$

and

$$\mathrm{cotan}^2(\hat{A}) = a^2$$

since

$$\cos^2\theta = \frac{\mathrm{cotan}^2\theta}{1 + \mathrm{cotan}^2\theta}$$

□

Homework (B6.1)**.** *Given $V_{p\times n}$, $U_{n\times 1}$, $\|U\| = 1$. and $A_{U,V}$ is the minimum angle between U and $L_{row}(V)$, show that*

$$\mathrm{cotan}^2(A_{U,V}) = S_{UV}\left(\overset{\perp}{S}_{VV}\right)^{-1} S_{VU}$$

where

$$S_{UV} = U'V'$$
$$S_{VU} = S'_{UV}$$
$$\overset{\perp}{S}_{VV} = V(I - UU')V'$$

Homework (B6.2)**.** *Apply HWB6.1 to the case of $p = 1$ and show that you get one-sample t-test.*

11.0.2. *Null-distribution.* (H_0 is true)

$$\hat{A} = \text{minimum angle between } L_{row}(\underset{\sim}{X}) \text{ and } \mathbf{1}_n$$
$$T^2 = (n-1)\text{cotan}^2(\hat{A})$$

The trick: Originally $L_{row}(\underset{\sim}{X})$ is "uniform" over all p-dimensional subspaces of $\mathbb{R}^n$, $\mathbf{1}_n$ fixed. Now, consider the case $L_{row}(\underset{\sim}{X})$ is fixed and replace $\mathbf{1}_n$ by a "Uniform" direction.

$$\underset{\sim}{X}' = \underbrace{\Gamma}_{n\times p} T$$

By this, we have $\hat{A}$ has the same distribution as the minimum angle between any realization of $L_{row}(\underset{\sim}{X})$ and $y/\|y\|$, $y \sim \mathcal{N}(\mathbf{0}, I_n)$.

Take $L_{row}(\underset{\sim}{X}) = L_{row}(e_1, e_2, \ldots, e_p)$ where e_i vector with 1 in its ith entry and 0 everywhere else.

Now,

$$\begin{aligned}\text{cotan}^2(\hat{A}) &= \frac{y_1^2 + \cdots + y_p^2}{y_{p+1}^2 + \cdots y_n^2} \\ &= \frac{p}{n-p}\left(\frac{(y_1^2 + \cdots + y_p^2)/p}{(y_{p+1}^2 + \cdots + y_n^2)/(n-p)}\right) \\ &= \frac{p}{n-p}F_{p,n-p}\end{aligned}$$

Now, $T^2 = (n-1)\text{cotan}^2(\hat{A}) = \frac{n-1}{n-p}pF_{p,n-p}$.

11.0.3. *Non-null T^2-distribution.* ($H_1 : \mu \neq 0$)

Invariance: Start with $\underbrace{\underset{\sim}{X}}_{p\times n} = \mathcal{N}_{p\times n}(\mu\mathbf{1}'_n, \Sigma_{p\times p} \bigotimes I_n)$ and A, which is $p \times p$ and nonsingular.

Let $\underset{\sim}{\tilde{X}}_{p\times n} = A\underset{\sim}{X}$. We have two versions of T^2. Applied to $\underset{\sim}{X}$, we have T^2 and applied to $\underset{\sim}{\tilde{X}}$, we have $\tilde{T}^2$.

$$\begin{aligned}\bar{X} &= \sum_{i=1}^n \frac{X_i}{n} \\ \bar{\tilde{X}} &= \frac{1}{n}\left(\sum_{i=1}^n AX_i\right) = A\bar{X} \\ S &= \underset{\sim}{X}\underset{\sim}{X}' - n\bar{X}\bar{X}' \\ \tilde{S} &\equiv \underset{\sim}{\tilde{X}}\underset{\sim}{\tilde{X}}' - n\underset{\sim}{\tilde{X}}\underset{\sim}{\tilde{X}}' \\ &= A[\underset{\sim}{X}\underset{\sim}{X}' - n\bar{X}\bar{X}']A' \\ &= ASA'\end{aligned}$$

Now,

$$\begin{aligned}\tilde{T}^2 &= \bar{\tilde{X}}'\left[\frac{\tilde{S}}{n(n-1)}\right]^{-1}\bar{\tilde{X}} \\ &= (A\bar{X})'\left[\frac{ASA'}{n(n-1)}\right]^{-1}A\bar{X} \\ &= \bar{X}'\left[\frac{S}{n(n-1)}\right]^{-1}\bar{X} \\ &= T^2\end{aligned}$$

This implies

$$L_{row}(\underset{\sim}{X}) = L_{row}(A\underset{\sim}{X})$$

so $\tilde{T}^2 = T^2$
Implication: Taking $A = \Sigma^{-1/2}$,

$$\underset{\sim}{X} \sim \mathcal{N}_{p\times n}(\mu \mathbf{1}_n', \Sigma \bigotimes I_n)$$
$$\underset{\sim}{\tilde{X}} \sim \mathcal{N}_{p\times n}(\tilde{\mu} \mathbf{1}_n', I_n \bigotimes I_n) \qquad (\tilde{\mu} = \Sigma^{-1/2}\mu)$$

Theorem 5 (Hotelling's theorem)**.** *If* $\underset{\sim}{X} \sim \mathcal{N}_{p\times n}(\mu \mathbf{1}_n', \Sigma \bigotimes I_n)$, $\Sigma > 0$, *then*

$$T^2 = \bar{X}' \left[\frac{S}{n(n-1)}\right]^{-1} \bar{X}$$

has distribution

$$T^2 = \frac{n-1}{n-p} p F_{p,n-p}(\underbrace{n\mu'\Sigma^{-1}\mu}_{n\|\tilde{\mu}\|^2})$$

where $F_{p,n-p}(\delta^2)$ *is a non-central* F *distribution.*

$$\frac{\chi^2_p(\delta^2)/p}{\chi^2_{n-p}(0)/(n-p)}$$

Lemma. *Let* U *be a fixed unit-norm* $p \times 1$ *vector,* $S \sim W(I_p, m, p)$ *with* $m \geq p$*. Then*

$$U'S^{-1}U \sim \frac{1}{\chi^2_{m-p+1}}$$

Proof. For any $\Delta_{p\times n}$ orthogonal matrix,

$$\begin{aligned} U'S^{-1}U &= U'\Delta'\Delta S^{-1}\Delta'\Delta U \\ &= (\Delta U)'(\Delta S^{-1}\Delta')(\Delta U) \\ &= \tilde{U}'\tilde{S}^{-1}\tilde{U} \end{aligned}$$

where

$$\begin{aligned} \tilde{S} &= \Delta S \Delta' \\ \tilde{U} &= \Delta U \end{aligned}$$

Then

$$\begin{aligned} \tilde{S} = \Delta' U \Delta &\sim W(\Delta\Delta'; m, p) \\ &= W(I_p; m, p) \end{aligned}$$

Implication: you can choose whatever Δ, orthogonal. In particular, we pick Δ such that

$$\Delta U = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

Then in this case,

$$\tilde{U} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

and

$$U'S^{-1}U = \tilde{U}\tilde{S}^{-1}\tilde{U} = (\tilde{S}^{-1})_{p,p}$$

Recall that

$$\begin{aligned} \tilde{S} &= \underset{\sim}{X}\underset{\sim}{X}' \\ &= T'\Gamma'\Gamma T \\ &= T'T \end{aligned}$$

where

$$\underset{\sim}{X}' = \Gamma_{n\times p} T_{p\times p} \qquad \underset{\sim}{X} \sim \mathcal{N}_{p\times n}(\tilde{\mu}, I_p \bigotimes I_n)$$

Now,

$$\tilde{S}^{-1} = (T'T)^{-1} = T^{-1}(T^{-1})'$$
$$(\tilde{S}^{-1})_{p,p} = (T^{-1}_{p,p})^2$$

Now what is $(T_{p,p})^{-1}$? Since $TT^{-1} = I_p$, and they are upper triangular,

$$T_{pp}(T^{-1})_{pp} = 1$$

□

Homework (B6.3)**.** *Finish the proof*

Proof of Hotelling's theorem. By B.1 (lecture),

$$\bar{X} \sim \mathcal{N}\left(\tilde{\mu}, \frac{I_p}{n}\right)$$

is independent of

$$S \sim W(I_p; n-1, p)$$

We also have

$$T^2 = \|\bar{X}\|^2 n(n-1) U' S^{-1} U \qquad U \equiv \frac{\bar{X}}{\|X\|}$$

Therefore,

$$T^2 | \bar{X} \sim \|\bar{X}\|^2 \frac{n(n-1)}{\chi^2_{n-p}}$$

This is independent of $\|\bar{X}\|^2 \sim \frac{1}{n}\chi^2_p(n\|\tilde{\mu}\|^2)$. This implies

$$\begin{aligned} T^2 &= \frac{p}{n-p} \frac{(n-1)n\|\bar{X}\|^2/p}{\chi^2_{n-p}/(n-p)} \\ &= \frac{(n-1)p}{n-p} F_{p,n-p}(n\|\tilde{\mu}\|^2) \end{aligned}$$

□

12. October 16th, 2013

12.1. 2-sample test.

- Univariace case $n = n_1 + n_2$.

$$\begin{aligned} x_1, \ldots, x_{n_1} &\sim \mathcal{N}(\mu_1, 1) \\ y_1, \ldots, y_{n_2} &\sim \mathcal{N}(\mu_2, 1) \end{aligned}$$

- Let $z = (x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}) \in \mathbb{R}^n$.
- Let $d = \left(-\frac{1}{n_1}, \ldots, -\frac{1}{n_1}, \frac{1}{n_2}, \ldots, \frac{1}{n_2}\right)$

Note:

$$\begin{aligned} d'z &= -\frac{\sum x_i}{n_1} + \frac{\sum y_i}{n_2} \\ &= \bar{y} - \bar{x} \end{aligned}$$

Note: $d'\mathbf{1}_n = 0$, $d \perp \mathbf{1}_n$

Definition 11.

$$T^2 = \frac{(\bar{y} - \bar{x})^2}{\left[\frac{n}{n_1+n_2}\frac{S_1+S_2}{n-2}\right]}$$

$$S_1 = \sum_{i=1}^{n_1} (x_1 - \bar{x})^2$$

$$S_2 = \sum_{i=1}^{n_2} (y_i - \bar{y})^2$$

Geometric interpretation:

$$\hat{Z}^{\perp} = Z - \hat{Z}$$

where $\hat{Z}$ is the projection to $L_{col}(\mathbf{1}_n)$.

Homework (B6.4)**.** *verify that* $t^2 = (n-2)\text{cotan}^2(\hat{A})$

$$\begin{aligned}\mathbb{E}\left[\overset{\perp}{Z}\right] &= \mathbb{E}[(z_1 - \bar{z}, z_2 - \bar{z}, \ldots, z_n - \bar{z})] \\ &= \left(\mu_1 - \frac{n_1\mu_1 + n_2\mu_2}{n}, \ldots, \mu_2 - \frac{n_1\mu_1 + n_2\mu_2}{n}\right) \\ &\propto (n_2(\mu_1 - \mu_2), \ldots, n_2(\mu_1 - \mu_2), n_1(\mu_2 - \mu_1)) \\ &\propto (-n_2, \ldots, -n_2, n_1, \ldots, n_1)\end{aligned}$$

Note: $\mathbb{E}\left[\overset{\perp}{Z}\right] \propto d$.
If $\mu_1 = \mu_2$, then $\overset{\perp}{\mu} = 0$.
Also note:

$$\begin{aligned}Z &= \mu + W \qquad & W &\sim \mathcal{N}_n(\mathbf{0}, I_n) \\ \overset{\perp}{Z} &= \overset{\perp}{\mu} + \overset{\perp}{W} \qquad & \overset{\perp}{W} &\sim \mathcal{N}_{n-1}(\mathbf{0}, I_{n-1}) \\ &\propto d\end{aligned}$$

12.2. **Hotelling's T^2 in two sample - p-dimensions.** If we have

$$\begin{aligned}X_1, \ldots, X_{n_1} &\sim \mathcal{N}(\mu_1, \Sigma) \\ Y_1, \ldots, Y_{n_2} &\sim \mathcal{N}(\mu_2, \Sigma)\end{aligned}$$

with null hypothesis

$$H_0 : \mu_1 = \mu_2$$

$$\text{Sufficient Statistics} \begin{cases} \bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \\ S_1 = \sum_{i=1}^{n_1}(X_i - \bar{X})(X_i - \bar{X})' \\ \bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i \\ S_2 = \sum_{i=1}^{n_2}(Y_i - \bar{Y})(Y_i - \bar{Y})' \end{cases}$$

and the T statisitic is given by

$$T^2 = (\bar{X} - \bar{Y})' \left[\frac{n}{n_1 n_2}\frac{S_1 + S_2}{n-2}\right]^{-1} (\bar{X} - \bar{Y})$$

Geometrically,

$$\underbrace{\underset{\sim}{Z}}_{p\times n} = [\underbrace{\underset{\sim}{X}}_{p\times n_1}, \underbrace{\underset{\sim}{Y}}_{p\times n_2}] = \begin{bmatrix} \cdots & z_1 & \cdots \\ \cdots & z_2 & \cdots \\ & \vdots & \\ \cdots & z_p & \cdots \end{bmatrix}$$

$$\overset{\perp}{\underset{\sim}{Z}} = \underset{\sim}{Z}P, \quad P = I_n - \frac{\mathbf{1}_n\mathbf{1}_n'}{2}$$

Then $\hat{A}$ becomes the smallest angle between d and all vectors in $L_{row}(\overset{\perp}{Z})$. This is the same as athe angle between d and $\hat{d}$, where $\hat{d}$ is the projection of d onto $L_{row}(\overset{\perp}{Z})$.

Claim.

$$T^2 = (n-2)\text{cotan}^2(\hat{A})$$

Homework (B6.5)**.** *Argue from our previous result that under the null hypothesis,*

$$T^2 \overset{H_0}{\sim} \frac{n-2}{n-p-1} pF_{p,n-p-1}$$

Homework (B6.6)**.** *"Prostate Cancer" data, which is roughly 6033 by 102. 50 of them are normal patients and the remaining 52 are cancer patients.*

(a) Compute the 2-sample T^2-statistic for

$$\underbrace{\underset{\sim}{Z}}_{10\times 102} = prostate(1:10)$$

which is the first 10 rows

(b) Repeat now, removing the second gene

(c) Compute individual two-sample t-test for the first 10 genes and say why (b) is very different from (a)

12.3. **Mahalanobis Distance.** The power of the T^2 test depends on

$$\Delta^2 = \mu' \Sigma^{-1} \mu$$

for a one-sample. For a two-sample, it depends on

$$\Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$$

This is defined as the Mahanobi distance between μ_1 and μ_2. Technically, this means that in the one-sample test, Δ^2 is the Mahalanobis distance between μ and 0.

12.3.1. *Relationship between T^2 and univariate test.* Let us go back to the one-sample case.

$$T^2 = \bar{X} \left[\frac{S}{n(n-1)} \right]^{-1} \bar{X} \quad (\star)$$

with

$$\underset{\sim}{X} \sim \mathcal{N}_{p \times n}(\mu \mathbf{1}_n', \Sigma \bigotimes I_n)$$

Let a be a fixed vector.

$$y = a' \underset{\sim}{X} = (a' X_1, \ldots, a' X_n)$$

then

$$a' X_i \sim \mathcal{N}(a' \mu, a' \Sigma a)$$

Ideally, we would want

$$\frac{a' \mu}{\sqrt{a' \Sigma a}}$$

to be as large as possible, so $a \propto \Sigma^{-1} \mu$.

But you do not know μ.

One-sample t^2 distribution:

$$t^2(a) = \frac{n \bar{y}^2}{\sum (y_i - y)^2 / (n-1)} = n(n-1) \frac{(a' \bar{X})^2}{a' S a} \quad (\star\star)$$

Connection: $\star$ and $\star\star$ are connected.

Homework (B6.7)**.** *Show that*

$$\max_a (t^2(a)) = T^2$$

where "maximum" is achieved at $\hat{a}$ where

$$\hat{a}' \underset{\sim}{X} \equiv \hat{V}$$

where $\hat{V}$ is the projection of $\mathbf{1}_n$ to $L_{row}(\underset{\sim}{X})$.

Why? When p is very large, say $p \leq n$.

$$T^2 \sim \frac{(n-1)p}{n-p} F_{p, n-p}$$

the power depends on $\mu' \Sigma^{-1} \mu$. In a microarray, μ would be very sparse, in the sense that most entries are 0.

12.4. **False Discovery Rate (FDR)-controlling Method.**

- For each gene, use two-sample t to compute a score .
- Select a small fraction of the genes such that FDR is small.

13. Review

Class 1

- Cauchy-Schwarz
- QR (Gram-Schmidt)

$$\underbrace{V}_{\substack{p \times q \\ q \leq p}} = \underbrace{Q}_{p \times q} \underbrace{R}_{q \times q}$$

  where R is upper triangular and V is full rank.
- $L_{col}(V) = L_{col}(Q)$
- More generally, multiplying a nonsingular matrix on the right does not change L_{col} and multiplying on the left does not change L_{row}
- Definition $\overset{\perp}{L}_{col}$
- rank
- Eigen-representation $G = \Gamma D \Gamma' = 1 = \sum_{i=1}^{r} d_j \gamma_j \gamma_j'$ where $\Gamma = [\gamma_1, \ldots, \gamma_r]$

Class 2

- $V = V_{n\times p} = (v_1, \ldots, v_p)$ of rank $p : p \leq n$. $\hat{\beta} = (V'V)^{-1}V'y$, $\hat{y} = V(V'V)^{-1}V'y$
- $\overset{\perp}{P} = I_n - p$
- $P = \Gamma\Gamma'$, $\Gamma = (\gamma_1, \ldots, \gamma_p)$
- $\|\overset{\perp}{y}\|^2 = \min_{v\in V} \|y - v\|^2$
- $\cos^2(\alpha_{y,\hat{y}}) \equiv \frac{\|\hat{y}\|^2}{\|y\|^2}$ where $\alpha_{y,\hat{y}}$ is $\min_{v\in V}(\alpha_{y,v})$.
- Application:
$$y, v_1, \ldots, v_p \in L^2(\Omega, \beta, P)$$
all with expectation 0. Then the linear projection is
$$\hat{y} = \sigma_{YV}\sigma_{VV}^{-1}\begin{pmatrix} v_1 \\ \vdots \\ v_p \end{pmatrix}$$
- $\hat{y}$ is uncorrelated with
$$\overset{\perp}{y} = y - \hat{y}$$
-
$$\frac{\mathrm{Var}(\hat{y})}{\mathrm{Var}(y)} = \frac{\sigma_{YV}\sigma_{VV}^{-1}\sigma_{VY}}{\sigma_{YY}} \equiv \rho^2_{Y|v_1,\ldots,v_p}$$
- Gram-Schmidt
$$V_{n\times p} = W_{n\times p}T_{p\times p}$$
where V is full rank and $p \leq n$.
$$w_1 = \frac{v_1}{\|v_1\|}$$
$$w_2 = \frac{\overset{\perp}{v}_2}{\|\overset{\perp}{v}_2\|}$$
$$\vdots$$
$$w_p = \frac{\overset{\perp}{v}_p}{\|\overset{\perp}{v}_p\|}$$
with
$$t_{ij} = w_i'v_j$$
$$t_{jj} = \|\overset{\perp}{v}_j\| > 0$$

Class 3

- $A = (a_1, \ldots, a_p)$ which is $p \times p$. The determinant $\pm|A|$ is the p-dimensional volume of the parallelepiped having its lower corner determined by $a_1, \ldots, a_p$.
- $\mathrm{tr}(XY) = \mathrm{tr}(YX)$ and $\mathrm{tr}(A) = \sum \lambda_i$ if A is symmetric
- SVD
- pseudo inverses

Class 4

- partitioned inverse
- Woodbury's Theorem
- Normal
- Normal density of n-samples.
- Wishart
$$\underbrace{\underset{\sim}{X}}_{p\times n} \sim \mathcal{N}_{p\times n}(\tilde{\mu}, \Sigma \bigotimes \Delta)$$
$$A\underset{\sim}{X}B \sim \mathcal{N}_{a,b}(A\mu B, A\Sigma A' \bigotimes B'\Delta B)$$
- Rotational Invariance
$$\underset{\sim}{X} = \begin{pmatrix} v_1' \\ \vdots \\ v_p' \end{pmatrix}$$
$v_1, \ldots, v_p$ though correlated having spherical symmetry. p -dimensional subspace $L_{col}(v_1, \ldots, v_p)$ is "uniformly"-distributed over all q-dimensional subspace of $\mathbb{R}^n$.

Let $\underset{\sim}{X} \sim \mathcal{N}_{p\times n}\left(\frac{\mu\cdot \mathbf{1}_n'}{\sqrt{n}}, I_p \bigotimes I_n\right)$ where Then

$$R(\hat{\mu}^{MLE} =$$

$$\underset{\sim}{X} = [x_1, \ldots, x_n$$

Let $Z = \frac{1}{\sqrt{n}}\sum_i X_i \sim \mathcal{N}(\mu, I_p)$.

14.1. **Stein's normal means problem.** Goal: Given $Z \sim \mathcal{N}(\mu, I_p)$, how do we estimate μ?

- Method 1: MLE

  $$f_Z(z) = \frac{1}{(\sqrt{2\pi})^p} e^{-\frac{1}{2}\|z-\mu\|^2}$$

  Then

  $$\hat{\mu}^{MLE} = Z$$

  The mean square error is

  $$MSE = R(\hat{\mu}, \mu) = \mathbb{E}\left[\sum_{i=1}^{p}(\hat{\mu}_i - \mu_i)^2\right]$$

  With the MLE, we have

  $$R(\hat{\mu}^{MLE}, \mu) = \mathbb{E}\left[\sum_{i}^{p}(z_i - \mu_i)^2\right] = p$$

- Method 2: Stein's shrinkage. (JS is James-Stein)

  $$\hat{\mu}^{JS} = \left(1 - \frac{p-2}{\|Z\|^2}\right) Z$$

  where

  $$R(\hat{\mu}^{JS}, \mu) < p$$

  for $p \geq 3$.

14.2. **Wiener Process/Image, Signal Processing/Non-parametric estimation.**

$$X(t) = f(t) + W(t)$$

where $0 \leq t \leq 1$ and $f(t)$ is a function over $[0,1]$ which is smooth, unknown, and or primary interest. $W(t)$ is a stationary Gaussian Process.

- Discretize

  $$t_0 = 0, t_1 = \frac{1}{p}, \ldots, t_p = \frac{p}{p} = 1$$

  with

  $$X(t_i) = f(t_i) + W(t_i)$$

  for $i = 0, 1, \ldots, p$.
- Use an orthogonal basis $\{\phi_0, \phi_1, \ldots, \phi_p, \ldots\}$ where the ϕ_i are unit l^2-norm functions. We take a finite number, say $p+1 = 2^N$. Take coefficients $c_0, c_1, \ldots, c_p$.

  $$\begin{aligned} Z_j &= \frac{1}{p}\sum_{i=0}^{p}\phi_j(t_i)X(t_i) \\ &\approx \int_0^1 \phi_j(t)X(t)\,\mathrm{d}t \\ &= \underbrace{\int_0^1 \phi_j(t)f(t)\,\mathrm{d}t}_{\mu_j} + \underbrace{\int_0^1 \phi_j(t)W(t)\,\mathrm{d}t}_{s_j} \end{aligned}$$

  Thus $Z_j = \mu_j + s_j$ and

  $$\boldsymbol{Z} \approx \mathcal{N}(\mu, I_p)$$

Sparsity:

$$\begin{aligned} |\mu_j| &= \left|\int_0^1 f(t)\phi_j(t)\,\mathrm{d}t\right| \\ &\leq j^{-\alpha} \end{aligned}$$

for $\alpha > 0$.

- Strict Sense:
$$S(\mu) = \{1 \le i \le p, \mu_i \neq 0\}$$
where $|S(\mu)| << p$.
- Loose Sense: "few" large coordinates and most coordinates are small.

If $Z \sim \mathcal{N}(\mu, I_p)$, we know that the MLE is inadmissible since Stein's shrinkage is better. However, this is also inadmissible since we can beat it. What do we do if μ is sparse?

$$\hat{\mu} = a(Z) \cdot Z$$

for $0 < a(Z) \le 1$, which is called a Linear estimator.

Hard Thresholding: Fix $t > 0$.

$$\hat{\mu}_i^{hs} = \begin{cases} Z_i & |Z_i| \ge t \\ 0 & \text{otherwise} \end{cases}$$

Soft Thresholding:

$$\hat{\mu}_i^{st} = \begin{cases} Z_i - t & Z_i \ge t \\ Z_i + t & Z_i \le -t \\ 0 & |Z_i| < t \end{cases}$$

then

$$\mu_i = \begin{cases} \ge \tau & \varepsilon\text{fraction} \\ Say 0 & (1-\varepsilon)\text{fraction} \end{cases}$$

Say, $\tau \ge \sqrt{2r \log p}$, $r > 4$.

Lemma. *with prob* $1 - O\left(\frac{1}{\log(p)}\right)^{1/2}$,

$$\max_{1 \le i \le p} \{|Z_i - \mu_i|\} \le \sqrt{2 \log p}$$

If we take $t = \sqrt{2 \log p}$

$$|\hat{\mu}_i| = |Z_i| \le \sqrt{2 \log p}$$

if $\mu_i = 0$.

$$|\hat{\mu}_i| = |\mu_i + Z_i| \ge \tau - \sqrt{2 \log p} > t$$

if $|\mu_i| \ge 0$, which implies $\hat{\mu}_i \neq 0$ if $\mu_i \neq 0$.

$$MSE \le c \log(p) \cdot p\varepsilon << p$$

Proof. Mill's Ratio:

$$\mathbb{P}\Big(\mathcal{N}(0,1) \ge \sqrt{2r \log p}\Big) \approx \frac{1}{\sqrt{2r \log p}} p^{-r}$$

□

15. October 23rd,2013

Example 10. *Consider* $Z \sim \mathcal{N}(\mu, I_p)$

$$\mu_i = \begin{cases} \tau & \textit{for } \varepsilon \textit{ fraction of } i, 1 \le i \le p \\ 0 & (1-\varepsilon) \end{cases}$$

Fix a t (may depend on p).

$$\hat{\mu}_i = \begin{cases} Z_i & |Z_i| \ge t \\ 0 & \textit{otherwise} \end{cases}$$

The mean square error is

$$\begin{aligned} MSE &= \mathbb{E}\Big[\sum (\hat{\mu}_i - \mu_i)^2\Big] \\ &= \sum_{i:\mu_i=0} \mathbb{E}\big[(\hat{\mu}_i - 0)^2\big] + \sum_{i:\mu_i=\tau} \mathbb{E}\big[(\hat{\mu}_i - \tau)^2\big] \\ &- \sum_{i:\mu_i=0} \mathbb{E}_0[Z_i^2 \mathbb{1}\{|Z_i| \ge t\}] \\ &- \sum_{i:\mu_i=\tau} \big(\mathbb{E}_1[\tau^2 \mathbb{1}\{|Z_i| \le t\} + \mathbb{E}_1[(Z_i - \tau)^2 \mathbb{1}\{|Z_i| \ge t\}\big) \\ &= p\varepsilon \end{aligned}$$

In summary,

$$MSE = p(1-\varepsilon)\mathbb{E}_0[Z_i^2 \mathbb{1}\{|Z_1| \geq t\} + p\varepsilon[\tau^2 P_0(|Z_i+\tau| \leq t) + \mathbb{E}_0[Z_i^2 \mathbb{1}\{|Z_i+\tau| > t\}]]$$

Homework (No number?)**.** *Let $t = t_p = \sqrt{2\log p}$, which is called the universal threshold and $\tau = \tau_p = \sqrt{2r\log p}$.*
Suppose $r > 1$ Show that

$$\begin{aligned} -P_0(|Z_i+\tau| \leq t) &= L_p p^{-(r-1)} \\ -\mathbb{E}_0[Z_i^2 \mathbb{1}\{|Z_i+\tau| \geq t\}] &= 1 + o(1) \\ -\mathbb{E}_0[Z_i^2 \mathbb{1}\{|Z_i| \geq t\}] &= L_p p^{-1} \end{aligned}$$

where L_p is the generic multi-log notation e.g. $(2\log)^{-3/2}\ (\log p)^{1/2}$

Answer:

$$\begin{aligned} \text{MSE} &\leq L_p \left[p(1-\varepsilon)\frac{1}{p} + p\varepsilon \left(1 + p^{-(r-1)}\right)\right] \\ &\leq L_p \cdot p \cdot \varepsilon \end{aligned}$$

Mill's ratio: $t > 0$. $E_1 : \mathcal{N}(\tau, 1)$

$$\left(\frac{1}{t} - \frac{1}{t^3}\right)\phi(t) < \frac{t}{1+t^2}\phi(t) \leq \int_t^\infty \phi(x)\,\mathrm{d}x \leq \frac{1}{t}\phi(t)$$

For E_0, the law of $\mathcal{N}(0,1)$

$$\left(t - \frac{1}{t}\right)\phi(t) \leq \int_t^\infty x^2\phi(x)\,\mathrm{d}x \leq t\phi(t)$$

Suppose $\mu_i \sim (1-\varepsilon)\nu_0 + \varepsilon \cdot F$

$$R^* = \inf_{\hat{\mu}} \sup_{(\varepsilon,F)} \text{MSE}(\hat{\mu}, \mu) \qquad (\star)$$

Under L^q-ball property, $\star$) $\mathbb{E}[|\mu_i|^q] \leq \eta^p$, $\eta = \eta_p \to 0$, $0 < q < 2$.
Given a prior $\varepsilon \cdot F \longrightarrow \hat{\mu}^B \longrightarrow \text{MSE}_{(\varepsilon,F)}(\hat{\mu}^B, \mu)$.
For some choice of (ε, τ), the prior

$$(1-\varepsilon)\nu_0 + \frac{\varepsilon}{2}\nu_\tau + \frac{\varepsilon}{2}\nu_{-\tau/2}$$

- FDR controlling method
- L^0/L^1 - penalization method

L^0: Fundamentally correct method for variable selection (in some cases) and why in modern regime (genetics), $L^)$ is not fundamentally correct.

$$l^1\text{-solution} \approx L^0\text{-solution} \approx \text{truth}$$

More realistically,

$$\begin{cases} \mu_i \neq 0 & \text{for,say } \varepsilon \text{ fraction of } i \\ \mu_i = 0 & (1-\varepsilon) \text{ fraction} \end{cases}$$

and

$$\hat{\mu}_i = \begin{cases} Z_i & |Z_i| \geq t \\ 0 & |Z_i| < t \end{cases}$$

when all $|\mu_i| \geq \tau$ for τ large, the previous argument says

$$\begin{aligned} \mu_i \neq 0 &\Leftrightarrow \hat{\mu}_i \neq 0 \\ \mu_i = 0 &\Leftrightarrow \hat{\mu}_i = 0 \end{aligned}$$

When τ is not that large, it's impossible to separate zero μ_i and nonzero μ_i. It also does not make sense to use the universal threshold $t = \sqrt{2\log p}$.

15.1. **Shift of Regime.** Strong signal to weak signals, τ large to τ "small", Engineering to genetic.

Goal: Go from "Completely" separate separate from noise to Identifying some coordinates, almost all of which are signals, though many signals are left behind.

Definition 12. *We call "i" a discovery if* $\hat{\mu}_i \neq 0$.

- True discovery: $\hat{\mu}_i \neq 0$, $\mu_i \neq 0$
- False discovery: $\hat{\mu}_0 \neq 0$, $\mu_i = 0$

The false discovery rate is

$$\text{FDR} = \mathbb{E}\left[\frac{\text{False Discoveries}}{\text{All Discoveries}}\right]$$

We hope to have a method such that the FDR is less than or equal to 5%.

15.2. **Benjamin-Hochberg's FDR Controlling Procedure.**

- Compute p different p-values
$$\pi_i = \mathbb{P}(|\mathcal{N}(0,1)| \geq |Z_i|)$$
- Sort them
$$\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$$
- Let $q \in (0,1)$ and $k = k^q_{FDR}$ be the largest integer such that
$$\pi_{(k)} \leq q\left(\frac{k}{p}\right)$$

Let t^q_{FDR} be the threshold such that

$$\mathbb{P}(|\mathcal{N}(0,1)| \geq t^q_{FDR}) = \pi_{(k^q_{FDR})}$$

e.g. if $k^q_{FDR} = 10$, then we take $\pi_{(10)}$. We then take

$$\hat{\mu}_i^{FDR} = \begin{cases} Z_i & \text{if } |Z_i| \geq t^q_{FDR} \\ 0 & \text{otherwise} \end{cases}$$

15.3. **Large-Scale Multiple Testing.** H_{0i}(null) vs H_{1i}(alternative), $i = 1, 2, \ldots, p$. We have a summary statistic associated with each i, say Z_i. In genomics, for example, we have

$$H_{0i} : \text{not differentially expressed}$$
$$H_{1i} : \text{differentially expressed}$$

$$\begin{aligned} T_i &= \text{studentized } t \\ &= \frac{\text{mean in one group} - \text{mean in other}}{\sqrt{\frac{S_1}{n_1} + \frac{S_2}{n_2}}} \end{aligned}$$

where n_1 is the number of subjects in group 1 (normal patients) and n_2 is the number of subjects in group 2 (diseased patients) and $i = 1, 2, \ldots, p$.

$$\pi_i = P_{H_{0i}}(|F_{0i}| \geq |Z_i|)$$

Theorem 6. *If* π_i*, associated with* (H_{01}, H_{1i}, Z_i)

- *Independent for all* $1 \leq i \leq p$.
- $(1-\varepsilon)$ *fraction of i:* $\pi_i \sim \mathcal{U}(0,1)$
- ε *fraction of i:* $\pi_i \sim F_i$

Then FDR by Benjamin-Hochberg is $q(1-\varepsilon) \leq q$.

16. October 28th, 2013

16.1. **Variable Selection.** We have

$$Y = X\beta + Z$$

where $Z \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$, $X = X_{n,p}$, and β a coefficient with some non-zero entries and some zero entries.

We have three parts

- 1970-1980, $n >> p$, AIC, BIC, Mallon's C_p, and forward/backward regression
- late 1980s "Wavelet and Image processing" $p > n$.

  Lion face: Skin - Fourier Basis. Hair - Dirac basis.

  $$X = [\Phi, \Psi]$$

  where Φ and Ψ are $n \times n$ orthogonal bases. β has $<< n$ non-zero entries.
  - You can choose X, therefore X is "nice".

  - β is sparse.
  - There is very little noise, so $\sigma = 0$ or "small".

  We can say that the off diagonals of $G = X'X$ are small.
- Genetics/genomics
  - β is sparse.
  - You cannot choose X, so X could be "bad".
  - σ "large".

  In this case, the off diagonals would be large of close to 1.
- "Signal Recovery" Problem

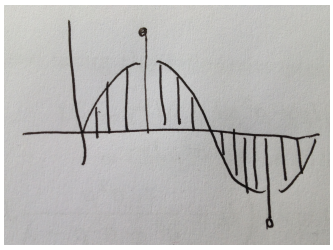
  We have that $Y = X\beta$ (i.e. $\sigma = 0$)

$$X = [\Phi, \Psi]$$

  where Φ and Ψ are $n \times n$, $p = 2n$. β is sparse:

$$\mathcal{S} \equiv S(\beta) = \{1 \leq i \leq p : \beta_i \neq 0\}$$

  and $s = |S| << n$.



$$\begin{aligned} Y_t &= \text{Spike}(t) + \text{Sinusoid}(t) \\ &= \sum_{t=0}^{n-1} X_t \delta_t + \sum_{t=0}^{n-1} W_k \text{Sinusoid}_k(t) \end{aligned}$$

Many $X_t = 0$, many $W_k = 0$. where

$$\delta_t(x) = \begin{cases} 1 & x = t \\ 0 & \text{otherwise} \end{cases}$$

Here,

$$\beta = \begin{pmatrix} (X_t)_{0 \leq t \leq n-1} \\ (W_k)_{0 \leq k \leq n-1} \end{pmatrix}$$

which is very sparse.

- A little bit of quantum mechanics. The state of a 1-d particle is described by a wave function, $f \in \mathcal{L}^2(\mathbb{R})$.
- The prob. density that it is at location t is

$$\frac{1}{\|f\|^2} |f(t)|^2$$

- The prob. density that its momentum is at ω is

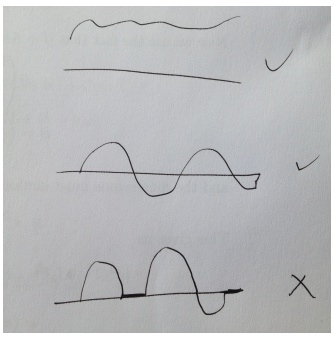
$$\frac{1}{2\pi \|f\|^2} |\hat{f}(\omega)|^2, \qquad \hat{f}(\omega) = \int f(t) e^{-itw\omega} \, \mathrm{d}t$$

  where $i = \sqrt{-1}$. By Parseval's inequality,

$$w\pi \|f\|^2 = \|\hat{f}\|^2$$

Theorem 7 (Heisenberg's Uncertainty Principle)**.** $\sigma_t^2 \sigma_\omega^2 \geq 1/4$

Theorem 8 (Heisenberg's Uncertainty Principle)**.** *If $f \neq 0$ has compact support, then $\hat{f}(\omega)$ cannot have 0 over a whole interval.*



(that is the set where $f = 0$ cannot have more than measure zero.

16.2. **Discrete Uncertainty Principle.** Let $(X_t)_{t=0,1,\ldots,n-1}$ be a length n sequence. Let $(\hat{X}_\omega)_{\omega=0,1,\ldots,n-1}$ be a discrete Fourier transformation.

- $$\hat{X}_\omega = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} X_t e^{-2\pi i \omega t/n} = \Psi^*(X_t)$$
  for $\omega = 0, 1, \ldots, n-1$.
- $\Phi = I_n$
- Ψ is such that
  $$\Psi_{jk} = \frac{1}{\sqrt{n}} e^{2\pi i jk/n}$$
  where $i = \sqrt{-1}$.

Theorem 9 (DUT)**.** *For any n-vector $x \neq 0$*

$$|T| + |W| \geq 2\sqrt{|T||W|} \geq 2\sqrt{n}$$

where

$$T = \{0 \leq t \leq n-1 : X_t \neq 0\}$$
$$W = \{0 \leq \omega \leq n-1 : \hat{X}_\omega \neq 0\}$$

Corollary. *We can not have two β, (β_1, β_2) such that*

$$\begin{cases} Y = X\beta_1 = X\beta_2 \\ \textit{Both } \beta_1, \beta_2 \textit{ satisfy } |T| + |W| < \sqrt{n} \end{cases}$$

$$\underbrace{\beta}_{2n \times 1} = \begin{pmatrix} \beta_T \\ \beta_W \end{pmatrix}$$

where β_T is $n \times 1$ and β_W is $n \times 1$ and T is the support of β_T and W is the support of β_W.

Proof of Corollary. If β_1 and β_2 exist, write $\alpha = \beta_1 - \beta_2$.

$$X\alpha = X\beta_1 - X\beta_2 = 0$$

Then

$$X = [\Phi, \Psi] \qquad \alpha = \begin{pmatrix} \alpha_T \\ \alpha_W \end{pmatrix}$$

and

$$\begin{aligned} X\alpha = 0 &\Leftrightarrow \alpha_T + \Psi\alpha_W = 0 \\ &\Leftrightarrow \alpha_W = -\Psi^*\alpha_T \end{aligned}$$

In other words, if

$$\alpha_T = (X_t)$$

for short, then

$$\alpha = \begin{pmatrix} (X_t) \\ -(\hat{X}_\omega) \end{pmatrix}$$

For β_1 and β_2, we have $|T| + |W| < \sqrt{n}$. Then for α,

$$|T| + |W| < 2\sqrt{n}$$

Use DUP, $\alpha = 0$, and $\beta_1 = \beta_2$. □

Interpretation: Consider equation (Y, X given).

$$Y = X\beta$$

If $p = 2n > n$, then we have infinitely many solutions. However, the corollary satys that only one (or none) solution satisfy $|T| + |W| < \sqrt{n}$. If the true β satisfies $|T| + |W| < \sqrt{n}$, then there is one but we cannot have another.

Surprise: While there are "many" solutions, only one could be sufficiently sparse ($|T| + |W| < \sqrt{n}$).

Occam Razor: If different models equally explain a phenomenon, we trust the simplest one.

Solve (P_0) which minimizes $\|\beta\|_0$ subject to $Y = X\beta$ and (P_0) which minimizes $\|\beta\|_1$ subject to $Y = X\beta$. This was done by Donoho and Starck in 1989.

Theorem 10. *Consider a linear model where*

$$|T| + |W| < \sqrt{n} \quad (\textit{true } \beta)$$

Then (P_0) *has a unique solution, coinciding with true* β_0.

Proof. Let β^* be the solution of (P_0) and let β be the truth.

$$\|\beta^*\|_0 \leq \|\beta\|$$

The goal is to show $\delta = 0$ if we denote $\delta = \beta^* - \beta$. Similarly, write $\delta = \begin{pmatrix} \delta_T \\ \delta_W \end{pmatrix}$. If $\delta \neq 0$, then

$$[\Phi, \Psi] \begin{pmatrix} \delta_T \\ \delta_W \end{pmatrix} = 0$$

is equivalent to

$$\begin{aligned} \delta_T &= -\Psi \delta_W \\ -\delta_W &= \Psi^* \delta_T \end{aligned}$$

which are the Fourier coefficients. Since β has $|T| + |W| < \sqrt{n}$,

$$\|\beta^*\|_0 \leq \|\beta\|_0$$

implies that β^* has less than $\sqrt{n}$ nonzeros, which implies δ has $< 2\sqrt{n}$ nonzeros. This is a contradiction for

$$\delta = \begin{pmatrix} \delta_T \\ -(\hat{\delta_T})_W \end{pmatrix}$$

□

Interpretation: If $Y = X\beta$ and β is sufficiently sparse in terms of $|T| + |W| < \sqrt{n}$, then
(a) Only one solution to $Y = X\beta$ can be this sparse
(b) (P_0) recovers this unique sparsest solution

Unfortunately, (P_0) is NP-hard.

(P_1): Interior Point Optimizer (1970).

1965: Logan BF (1965) Properties of high-pass signals.

By Tibshirani in 1996, the lasso is

- $\|Y - X\beta\|^2 + \lambda\|\beta\|_1$, $\lambda > 0$
- $|T| + |W| < \sqrt{n}/2$

A different look at DUP: Fix (T, W) (subsets of $\{1, 2, \ldots, n\}$). Define the "maximum fraction of concentration of l^0-norm" by

$$\mu_0(T, W) = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \left\{ \frac{\sum_{t \in T} |X_t|_0 + \sum_{\omega \in W} |\hat{X}_\omega|_0}{\|X\|_0 + \|\hat{X}_\omega\|_0} \right\}$$

The numerator is $\leq |T| + |W|$ and the denominator $\geq 2\sqrt{n}$. This implies that

$$\mu_0(T, W) \leq \frac{|T| + |W|}{2\sqrt{n}}$$

Interpretation: We can re-state DUP as follows: For any (T, W) and for any vector (X_t) and $(\hat{X}_\omega)$, the concentration of l^0-norm in (T, W)

$$\leq \mu_0(T, W) \leq \frac{|T| + |W|}{2\sqrt{n}}$$

Interpretation: Consider $Y = X\beta$ where β is so sparse that

$$|T| + |W| < \sqrt{n}$$

then $\mu_0(T, W)$ passes the critical threshold of Y_2

$$\mu_0(T, W) < \frac{1}{2}$$

and (P_0) has a unique solution, which is the true β.

Now, define the "maximum fraction of concentration of l^1-norm" by

$$\mu_1(T, W) = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \left\{ \frac{\sum_{t \in T} |X_t|_1 + \sum_{\omega \in W} |\hat{X}_\omega|_1}{\|X\|_1 + \|\hat{X}_\omega\|_1} \right\}$$

Theorem 11 (Uncertainty Principle (l^1-version))**.**

$$\mu_1(T,W) \leq \frac{|T|+|W|}{\sqrt{n}+1}$$

If $|T|+|W| \leq \sqrt{n}/2$, then $\mu_1(T,W) < 1/2$.

17. October 30th, 2013

Theorem 12 (Up, l^1-version)**.**

$$\mu_1(T,W) \leq \frac{|T|+|W|}{\sqrt{n}+1}$$

For any vector $\begin{pmatrix} X_T \\ \hat{X}_\omega \end{pmatrix}$*, you try to put/concentrate as more as* l^1*-norm in sets* $T \cup W$*. Then the fraction of concentration in* l^1*-norm*

$$\leq \frac{|T|+|W|}{\sqrt{n}+1}$$

and

$$X = \begin{pmatrix} X_T \\ \hat{X}_\omega \end{pmatrix}$$

Implication: If $|T|+|W| \leq \sqrt{n}/2$ then $\mu_1(T,W) < 1/2$.

Theorem 13. *Consider* $y = [\Phi, \Psi]\beta$ *where* $|T|+|W| \leq \sqrt{n}/2$*, with* $|T|+|W|$ *associated with* β*. Then* (P_1) *has a unique solution, coinciding with the solution of* P_0 *and the true* β*.*

Proof. (P_1): $\min \|\beta\|_1$ such that $Y = X\beta$. Let β be the truth and β^* be the solution of (P_1). Then $\|\beta^*\|_1 \leq \|\beta\|_1$. Let

$$\delta = \beta^* - \beta = \begin{pmatrix} \delta_T \\ -(\hat{\delta}_T)_W \end{pmatrix}$$

Then $X\delta = 0$

We show the proof by contradiction. To do so, we show that if $\delta \neq 0$, then

$$\|\beta^*\|_1 + \|\beta+\delta\|_1 > \|\beta\|_1 \qquad (\star)$$

To show $(\star)$, note

$$\delta = \begin{pmatrix} (X_t) \\ -(\hat{X})_W \end{pmatrix}$$

and $(X_t) = \delta_T$. By

$$\mu_1(T,W) \leq \frac{|T|+|W|}{\sqrt{n}+1} < \frac{1}{2}$$

By definition of $\mu_1(T,W)$, (apply definition to δ)

$$\frac{\sum_{\gamma\in(T\cup W)} |\delta_\gamma|}{\sum_{\gamma\in(T\cup W)} |\delta_\gamma| + \sum_{\gamma\in(T\cup W)^c} |\delta_\gamma|} < \frac{1}{2}$$

Subtract both sides by $\frac{1}{2}$

$$\frac{1}{2}\frac{\sum\limits_{\gamma\in(T\cup W)^c} |\delta_\gamma| - \sum\limits_{\gamma\in(T\cup W)} |\delta_\gamma|}{\sum\limits_{\gamma\in(T\cup W)^c} |\delta_\gamma| + \sum\limits_{\gamma\in(T\cup W)} |\delta_\gamma|} > 0$$

which implies

$$\sum_{\gamma\in(T\cup W)^c} |\delta_\gamma| > \sum_{\gamma\in(T\cup W)} |\delta_\gamma|$$

Now, $|x+y|-|x| \geq -|y|$

$$\|\beta+\delta\|_1 - \|\beta\|_1 = \sum_{\gamma\in(T\cup W)^c} |\delta_\gamma| + \sum_{\gamma\in(T\cup W)} |\beta_\gamma+\delta_\gamma| - |\beta_\gamma|$$

$T \cup W$ is the support of β, so when $\gamma \in (T\cup W)^c$, $\beta_\gamma = 0$. So

$$\|\beta+\delta\|_1 - \|\beta\|_1 \geq \sum_{\gamma\in(T\cup W)^c} |\delta_\gamma| - \sum_{\gamma\in(T\cup W)} |\delta_\gamma| > 0$$

which implies

$$\|\beta^*\|_1 = \|\beta+\delta\|_1 > \|\beta\|_1$$

so we have a contradiction. □

The argument is as follows. Given $\beta : Y = X\beta$. This gives us $T(\beta)$ and $W(\beta)$. Apply $\mu_1(T(\beta), W(\beta))$ to any vectors of the form $\begin{pmatrix} (X)_t \\ (\hat{X})_W) \end{pmatrix}$, including $\delta = \beta^* - \beta$.

Incoherence:

$$Y = [\Phi, \Psi]\beta$$

where Φ and Ψ are two general orthogonal bases. Then

$$M(\Phi, \Psi) = \max\left\{ \max_{i,j}\left\{|(\Phi^{-1}\Psi)_{ij}|\right\}, \max_{i,j}\left\{|(\Psi^{-1}\Phi)_{ij}|\right\} \right\}$$

$$M(I_n, \text{Fourier}) = M(\psi, \psi^*) = \frac{1}{\sqrt{n}}$$

Uncertainty principle for general ψ and Φ.

Theorem 14. *For any vectors α and β such that*

$$\Phi\alpha = \psi\beta$$

we have $|T||W| \geq m^{-2}$ (and so $|T| + |W| \geq 2\sqrt{|T||W|} \geq 2m^{-1}$.) where T and W are nonzero sites of α and β.

Theorem 15. *When*

$$\|\beta\|_0 \leq \frac{\sqrt{2} - 1/2}{M(\Phi, \psi)} \approx 0.9\sqrt{n}$$

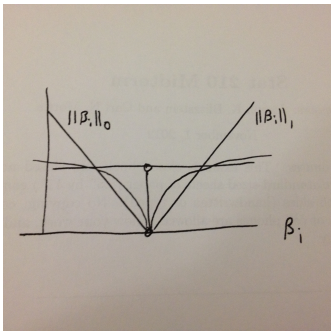
solutions to (P_0) and (P_1) are unique and the same as the truth.

Lasso: $Y = X\beta + Z$, $Z \sim \mathcal{N}(0, \sigma^2 I_n)$ with X arbitrary.

- $\|Y - X\beta\|_2^2 + \lambda\|\beta\|_0$
- $\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$
- SCAD Fan & Li 1998. MC+ Zhang C-H 2006

For β_i,

$$\|\beta_i\|_0 = \begin{cases} 1 & \beta_i \neq 0 \\ 0 & \beta_i = 0 \end{cases}$$



What could go wrong?

- Proved Facts
  - $\sigma = 0$ (noiseless)
  - X is nice, $X = [\Phi, \psi]$, $M(\Phi, \psi)$ small
  - $\|\beta\|_0 \leq \dfrac{c}{M(\Phi, \psi)}$
- Beyond this,
  (1) Whether it is fundamentally correct or not, Lasso/SCAD/MC+, you can always use
  (2) l^1-solution $\approx$ l^0-solution and SCAD/MC+ $\approx$ truth.

σ "large" and X is "bad" implies ???

Alternative:

- One step $\rightarrow$ multi-stage
- Global $\rightarrow$ local

C.1 Principle Component Analysis. Suppose $A_{p\times p} \geq 0$, symmetric, has spectral decomposition

$$\underbrace{A}_{p\times p} = \underbrace{\alpha}_{p\times p} D\alpha' = \begin{bmatrix} \vdots & & \vdots \\ \alpha_1 & \cdots & \alpha_p \\ \vdots & & \vdots \end{bmatrix} \begin{bmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_p \end{bmatrix} \begin{bmatrix} \cdots & \alpha_1' & \cdots \\ & \vdots & \\ \cdots & \alpha_p' & \cdots \end{bmatrix}$$

with $a_1 \geq a_2 \geq \cdots \geq a_p \geq 0$.

$$\gamma = \text{rank}(A) \leq p$$

We have $p - \gamma$ of the a_i are zero.

Quadratic Form:
For $g \in \mathbb{R}^p$,

$$\begin{aligned} Q(g) &= \|g\|_A^2 \\ &= g'Ag \\ &= \sum_{i=1}^{p} g' \left(a_i \alpha_i \alpha_i'\right) g \\ &= \sum_{i=1}^{p} a_i (\alpha_i' g)^2 \\ &= \sum_{i=1}^{p} a_i h_i^2 \end{aligned}$$

Ellipsoid: $\|g\|_A^2 = c$ determines a curve ellipsoid. h_i: $\alpha_i' g$ is a number. $B : \|g\|_B^2 \leq \|g\|_a^2$, $A \geq B$ ($A - B \geq 0$). If we don't have $A \geq B$ or $B \geq A$, then 2 ellipsoids would intersect.

ellipsoiddiagram

18. November 4th, 2013

We have that $A \geq B$ means that $A - B$ is positive semi-definite.
Ratio: Take $B = I_p$ (not assuming $A \geq B$).

$$\|g\|_B^2 = \|g\|^2 = \sum g_i^2$$

Then

$$R(g) = \frac{\|g\|_A^2}{\|g\|^2}$$

$$\begin{aligned} R(cg) &= \frac{\|cg\|_A^2}{\|cg\|^2} \\ &= \frac{\|g\|_A^2}{\|g\|^2} \\ &= R(g) \end{aligned}$$

so the ratio $R(g)$ is scaling invariant. We can always consider $\|g\|_A^2$ up to $\|g\| = 1$.

Lemma (Fundamental Lemma)**.** *Assume that all the $a_i > 0$ for convenience.*

(1) $g = \alpha_1$ maximizes $\|g\|_A^2$ (and so $R(g)$) subject to $\|g\|^2 = 1$ with maximum value $d_1 = \|\alpha_1\|_A^2$
(2) Among all vectors g satisfying

$$\begin{cases} g'\alpha_1 = 0 \\ \|g\|^2 = 1 \end{cases}$$

the one the maximizes $\|g\|_A^2$ is α_2 with maximum value $a_2 = \|\alpha_2\|_A^2$
(i) Among all vectors g satisfying

$$\begin{cases} g'\alpha_1 g'\alpha_2 = \cdots g'\alpha_{i-1} = 0 \\ \|g\|^2 = 1 \end{cases}$$

α_i maximizes $\|g\|_A^2$, with maximum a_i.
(p) $g = \alpha_p$ minimizes $\|g\|_A^2$, and so $R(g)$ on $\|g\|^2 = 1$ with value a_p.

Proof. The mapping

$$\underbrace{h}_{p\times 1} = \underbrace{\alpha'}_{p\times p}\underbrace{g}_{p\times 1}$$

with $g = \alpha h$ rotates coordinates. Take $\alpha_i \to e_i$.

$$\alpha'\alpha_i = \begin{pmatrix} \cdots & \alpha_1' & \cdots \\ & \vdots & \\ \cdots & \alpha_p' & \cdots \end{pmatrix} \alpha_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Now, $\|g\|^2 = \|h\|^2$.

$$\begin{aligned} \|g\|_A^2 &= g'Ag \\ &= g'\alpha D_a \alpha' g \\ &= h' D_a h \\ &= \|h\|_{D_a}^2 \end{aligned}$$

Then if $h = e_i$, then $g = \alpha e_i = \alpha_i$. Then we have to find the h such that $\|h\| = 1$ which maximizes

$$\sum_{i=2}^{p} a_i h_i^2$$

□

Homework (C1.1)**.** *Prove (p)*

Remark: Orthogonality in two matrices. For any vector g,

$$g'A\alpha_i = g'$$

Since $A = \alpha D \alpha'$, we have

$$\begin{aligned} g'A\alpha_i &= g'\left[\sum_{j=1}^{p} a_j \alpha_j \alpha_j'\right]\alpha_i \\ &= g' \sum_{j=1}^{p} a_j (a_i, a_j) a_j \\ &= a_i g' \alpha_i \end{aligned}$$

as we had that

$$(a_i, a_j) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

What this implies is that

$$\langle g, \alpha_i \rangle_A = a_i \langle g, \alpha_i \rangle$$

Therefore,

$$\langle g, \alpha_i \rangle_A = 0 \Leftrightarrow \langle g, \alpha_i \rangle = 0$$

18.0.1. *Principle Component in Sample Spaces.*

- Data Matrix

$$\underbrace{\underset{\sim}{X}}_{p\times n} = [X_1, \ldots, X_n] = \begin{bmatrix} \cdots & v_1' & \cdots \\ \cdots & v_2' & \cdots \\ & \vdots & \\ \cdots & v_p' & \cdots \end{bmatrix}$$

- Usually, we remove the means
- "Principle Components": a technical way for summarizing $\underset{\sim}{X}$ more succinctly.

Let $S_{p\times p} = \underset{\sim}{X}\underset{\sim}{X}' = LDL' = \sum_{i=1}^{p} d_i l_i l_i'$ where $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ are the eigenvalues and $l_1, \ldots, l_p$ are the corresponding eigenvectors.

$$\widehat{\text{Cov}} = \frac{S}{n}$$

Notation: $L_{(j)}$ is a $p \times j$ matrix with

$$L_{(j)} = [l_1, l_2, \ldots, l_j]$$

and

$$D_{(j)} = \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_j \end{bmatrix}$$

is a $j \times j$ matrix.

Define

$$Y_{(j)} = L_{(j)}'\underset{\sim}{X} = \begin{bmatrix} \cdots & Y_1 & \cdots \\ \cdots & Y_2 & \cdots \\ & \vdots & \\ \cdots & Y_j & \cdots \end{bmatrix} = \begin{bmatrix} l_1'\underset{\sim}{X} \\ l_2'\underset{\sim}{X} \\ \vdots \\ l_j'\underset{\sim}{X} \end{bmatrix} = \begin{bmatrix} l_1'(X_1, \ldots, X_n) \\ l_2'(X_1, \ldots, X_n) \\ \vdots \\ l_j'(X_1, \ldots, X_n) \end{bmatrix}$$

Each principle component is an n-dimensional vector that summarizes something for each student.

$Y_{(j)}$ is the j-th principle component representation of $\underset{\sim}{X}$ (rank j). The i-th component of Y_j is sometimes called the "loading" of X_i on Y_j.

$$Y_{(j)i} = l_j' x_i$$

Homework (C1.7)**.** *Redo all 3 figures for the score data. [Note, we have normalized the rows].*

Example 11 (Prostate Cancer Data)**.** *We have a* 6033×102 *matrix.* ($p = 6033, n = 102$) *and 5 of these are the controls and the remaining 52 are have the disease. Here* $p >> n$.

Two doubly standardized matrices $X1$ and $X2$ are such that $X1$ is 6033×50 and $X2$ is 6033×52.

$$\widehat{\text{Cov}}1 \text{ is } 50 \times 50 \qquad \widehat{\text{Cov}}2 \text{ is } 52 \times 52$$

In the figure, we have l_1 for $\widehat{\text{Cov}}1$ and l_1 for $\widehat{\text{Cov}}2$.

Can we do PCA on this? We need a very strong signal and l_1 such that l_1 is close to the l_1 of the underlying $\Sigma_{p,p}$.

$$\underbrace{\underset{\sim}{X}}_{p\times n} = \mu_0 \mathbf{1}_n' + \mu l + \underset{\sim}{Z}$$

where μ are the contrast means, l are the labels and $\underset{\sim}{Z}$ are the noise components.

Theorem 16. *Among all p-dimensional unit vectors* l, $\|l\| = 1$, *the first eigenvector* l_1 *maximizes*

$$\sum_{i=1}^{n} (l' x_i)^2 = \|l'\underset{\sim}{X}\|^2$$

with maximum d_1.

$$S = \underset{\sim}{X}\underset{\sim}{X}' = (l_1, \ldots, l_p) \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_j \end{pmatrix} \begin{pmatrix} l_1' \\ \vdots \\ l_p' \end{pmatrix}$$

Among al unit vectors l *with*

$$\begin{cases} l' l_1 = 0 \\ \|k\| = 1 \end{cases}$$

$l = l_2$ *maximizes* $\|l'\underset{\sim}{X}\|^2$

Homework (C1.2)**.** *Verify the theorem*

The full set of principle components

$$\underbrace{\underset{\sim}{Y}}_{p\times n} = \underbrace{L'}_{p\times p}\underbrace{\underset{\sim}{X}}_{p\times n} \equiv \begin{bmatrix} \cdots & Y_1 & \cdots \\ \cdots & Y_2 & \cdots \\ & \vdots & \\ \cdots & Y_p & \cdots \end{bmatrix}$$

has inner product matrix

$$\begin{aligned} \underset{\sim}{Y}\underset{\sim}{Y}' &= L'\underset{\sim}{X}\underset{\sim}{X}'L \\ &= L'SL \\ &= L'(LDL')L \\ &= D \end{aligned}$$

Homework (C1.3)**.** *Show that*

$$\sum_{k=1}^{j}\sum_{i=1}^{n} \underset{\sim}{Y_{(j)}}_{ki}^{2} = \sum_{k=1}^{j} d_k$$

where $Y_{(j)}$ is a $j \times n$ matrix. So $\underset{\sim}{Y}_{(j)}$ explains

$$\frac{\sum_1^j d_k}{\sum_1^p d_k}$$

proportion of

$$\sum_{k=1}^{p}\sum_{i=1}^{n} X_{ki}^2$$

19. November 6th, 2013

Now we do PCA on a random vector X, which is $p \times 1$ and look at interpretation.
Suppose X, which is $p \times 1$ is a member of $L^2(\Omega, \beta, P)$.

- $$X \sim (\underbrace{\mu}_{p\times 1}, \underbrace{\Sigma}_{p\times p})$$
- Subtracting off μ, makes $X \sim (0, \Sigma)$.
- Suppose $\Sigma_{p\times p} = \Gamma\Lambda\Gamma' = \sum_{i=1}^{p} \lambda_i\gamma_i\gamma_i'$ with $\Gamma = [\gamma_1, \ldots, \gamma_p]$, $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Theorem 17. *$\gamma = \gamma_1$ maximizes $\operatorname{Var}(\gamma' X)$, subject to $\|\gamma\| = 1$ with maximum λ_1; among all unit vectors satisfying $\gamma'\gamma_1 = 0$ (or equivalently, $\gamma' X$ and $\gamma_1' X$ are uncorrelated).*
$\gamma = \gamma_2$ maximizes $\operatorname{Var}(\gamma' X)$, maximum $= \lambda_2$.

Proof. Fundamental Lemma: if we take inner product to be

$$\gamma_1'\Sigma\gamma = \operatorname{Cov}(\gamma_1' X, \gamma' X)$$

□

Homework (C1.4)**.** *Finish the proof*

The random vectors

$$y_i = \gamma_i' X, i = 1, 2, \ldots, p$$

are the 1st, 2nd, …, pth principle components of the distribution of X. Let

$$Y_{p\times 1} = \Gamma'_{p\times p} X_{p\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}$$

$$\operatorname{Cov}(Y) = \Gamma'\Sigma\Gamma = \Gamma'\Gamma\Lambda\Gamma'\Gamma = \Lambda$$

Therefore, Y_i's are uncorrelated with variance λ_i.

The reason: dimension reduction

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{p} \lambda_i} \approx 0.99$$

Homework (C1.5)**.** *Show*

$$\operatorname{tr}(\Sigma) = \sum_{i=1}^{p} \lambda_i = \mathbb{E}\left[\|X - \mu\|^2\right] = \sum_{i=1}^{p} \operatorname{Var}(X_i) = \sum_{i=1}^{p} \operatorname{Var}(Y_i)$$

Homework (C1.6)**.** *Suppose $\mu = 0$, all $\lambda_i > 0$, and let*

$$\Gamma_{(j)} = (\gamma_1, \ldots, \gamma_j), j \leq p$$

(a) Show that the best linear predictor of X, in terms of

$$\underbrace{Y_{(j)}}_{j\times 1} = \Gamma'_{(j)} X$$

is

$$\underbrace{\hat{X}}_{p\times 1} = \sum_{i=1}^{j} \gamma_i y_i$$

(b) The residual $\overset{\perp}{X}_{p\times 1} = X - \hat{X}$ has covariance matrix

$$\overset{\perp}{\Sigma}_{(j)} = \sum_{i=j+1}^{p} \lambda_i \gamma_i \gamma_i'$$

with

$$\mathrm{tr}(\overset{\perp}{\Sigma}_{(j)}) = \sum_{i=j+1}^{p} \lambda_i$$

(c) For any matrix A, which is $j \times p$, let $Z = AX$ and

$$\overset{\perp}{X}_j = X - \underbrace{\Sigma_{XZ}\Sigma_{ZZ}^{-1} Z}_{best\ linear\ predictor}$$

Show that

$$\mathrm{tr}(\overset{\perp}{\Sigma}_Z) \equiv \mathrm{Cov}(\overset{\perp}{X}_Z) \geq \sum_{i=j+1}^{p} \lambda_i$$

Problem: For score data, $p < n$, it is reasonable to use PCA. For prostate data, $p >> n$.

19.1. **Definitions.**

$$\underbrace{\underset{\sim}{X}}_{p\times n} = (X_1, \ldots, X_n) = \begin{pmatrix} \cdots & v_1' & \cdots \\ \cdots & v_2' & \cdots \\ & \vdots & \\ \cdots & v_p' & \cdots \end{pmatrix}$$

SVD:

$$X_{p\times n} = L_{p\times r} C_{r\times r} R'_{r\times n}$$

$L = [l_1, \ldots, l_r]$, $C = \mathrm{diag}(c_1, \ldots, c_r)$, $R = [r_1, \ldots, r_r]$ with $c_1 \geq c_2 \geq \cdots \geq c_r > 0$.

$$L_{col}(\underset{\sim}{X}) = L_{col}(L), \qquad L_{row}(\underset{\sim}{X}) = L_{row}(R)$$

Then we have $S_{p\times p} = \underset{\sim}{X}\underset{\sim}{X}' = (LCR')(RCL') = LC^2L'$ with eigenvalues c_i^2 and eigenvectors l_i. Notation:

$$\underbrace{L_{(j)}}_{p\times j} = [l_1, l_2, \ldots, l_j], \qquad \underbrace{C_{(j)}}_{j\times j} = \mathrm{diag}(c_1, \ldots, c_j), j \leq r$$

and

$$R_{(j)} = (r_1, r_2, \ldots, r_j)$$

Lemma.

$$\begin{aligned} L_{(j)} L'_{(j)} \underset{\sim}{X} &= \underset{\sim}{X} R_{(j)} R'_{(j)} \\ &= L_{(j)} C_{(j)} R'_{(j)} \\ &\equiv \underbrace{\hat{X}}_{p\times n} \end{aligned}$$

$$\hat{X}_{(j)} = [\hat{X}_{1(j)}, \hat{X}_{2(j)}, \ldots, \hat{X}_{n(j)}] = \begin{bmatrix} \cdots & v'_{1(j)} & \cdots \\ \cdots & v'_{2(j)} & \cdots \\ & \vdots & \\ \cdots & v'_{p(j)} & \cdots \end{bmatrix}$$

Homework (C2.1)**.** *Verify the lemma.*

$\hat{\underset{\sim}{X}}_{(j)}$ has ith column

$$\hat{X}_{i(j)} = \underbrace{L_{(j)}}_{p\times j}\underbrace{L'_{(j)}}_{j\times p}\underbrace{X_i}_{p\times 1}$$

is the projection of X_i to $L(l_1, l_2, \ldots, l_j)$.
Residual:

$$\overset{\perp}{X}_{i(j)} = X_i - \hat{X}_{i(j)} = (I_p - L_{(j)}L'_{(j)})X_i$$

where $L_{(j)}L'_{(j)}$ is the projection to $\overset{\perp}{L}(l_1, \ldots, l_j)$.
Matrix norm: For $A_{p\times n}$ and $B_{p\times n}$ matrices, define the inner product

$$\langle A, B\rangle = \mathrm{tr}(AB') = \sum_{i=1}^{p}\sum_{k=1}^{n} a_{ik}b_{ik}$$

giving Frobenius norm

$$\|A\|^2 = \langle A, A\rangle = \sum_{i=1}^{p}\sum_{k=1}^{n} a_{ik}^2 = \mathrm{tr}(AA')$$

Fact 1: $\hat{\underset{\sim}{X}}_{(j)}$ and $\overset{\perp}{\underset{\sim}{X}}_{(j)} = \underset{\sim}{X} - \hat{X}_{(j)}$ are orthogonal with $\|\hat{\underset{\sim}{X}}_{(j)}\|^2 = \sum_{h=1}^{j} c_h^2$ and $\|\overset{\perp}{\underset{\sim}{X}}_{(j)}\|^2 = \sum_{h=j+1}^{r} c_h^2$, which implies

$$\|\hat{X}_{(j)}\|^2 + \|\overset{\perp}{X}_{(j)}\|^2 = \|X\|^2 - \sum_{h=1}^{r} c_h^2$$

Proof. By definition,

$$\begin{aligned}
\|\hat{\underset{\sim}{X}}_{(j)}\|^2 &= \mathrm{tr}(\hat{\underset{\sim}{X}}_{(j)}\hat{\underset{\sim}{X}}_{(j)}{}') \\
&= \mathrm{tr}[L_{(j)}C_{(j)}R'_{(j)}R_{(j)}C_{(j)}L'_{(j)}] \\
&= \mathrm{tr}(L_{(j)}C_{(j)}^2L'_{(j)}) \\
&= \mathrm{tr}(L'_{(j)}L_{(j)}C_{(j)}^2) \\
&= \sum_{h=1}^{2} c_h^2
\end{aligned}$$

□

Homework (C2.2)**.** *Finish the proof*

Theorem 18. *Among all j-dimensional subspaces of $\mathbb{R}^p$, $L(l_1, l_2, \ldots, l_j)$ maximizes the projected squares and minimizes the total orthogonal squared residuals.*

For any $P_j : \mathbb{R}^p \to$a 5-dimensional subspace,

$$\underset{\sim}{X} \to (PX_1, PX_2, \ldots, PX_n)$$

Loadings: From

$$\hat{\underset{\sim}{X}}_{(j)} = L_{(j)}[C_{(j)}R'_{(j)}]$$

we get ith column

$$X_{i(j)} = \sum_{h=1}^{j} l_h c_h r_{hi} \qquad (\star)$$

The notation is given by

$$R_{(j)} = [r_1, \ldots, r_j]$$

and

$$R'_{(j)} = \begin{bmatrix} \cdots & r'_1 & \cdots \\ \cdots & r'_2 & \cdots \\ & \vdots & \\ \cdots & r'_j & \cdots \end{bmatrix}$$

Fact 2: The coefficients of l_n in the expression of $\hat{X}_{i(j)}$ are the weighted coordinates of the ith column of $R'_{(j)}$ weighs c_h.
What's the role of L and R (and also C)?

- The first j left eigenvectors $l_1, \ldots, l_j$ determine the optimal j-dimensional projection space of $\mathbb{R}^p$.
- The right eigenvectors $r_1, \ldots, r_j$, weighted by $c_1, \ldots, c_j$ determine each X_i's projected coordinates.

Define

$$\underset{\sim}{Y}_{r\times n} = L_{r\times p}\underset{\sim}{X} = (Y_1,\ldots,Y_n) \equiv \begin{bmatrix} \cdots & r_1' & \cdots \\ \cdots & r_2' & \cdots \\ & \vdots & \\ \cdots & r_r' & \cdots \end{bmatrix} = L'LCR' = CR'$$

Compare this with $\star$.
Fact 3:

$$\hat{X}_{i(j)} - \sum_{h=1}^{j} l_h y_{h_i}$$

where $c_h r_{hi} \equiv y_{hi}$. The "loadings" y_{hi} are orthogonal in the sense

$$\underset{\sim}{Y}\underset{\sim}{Y}' = L'\underset{\sim}{X}\underset{\sim}{X}'L = L'LCR'RCL'L = C^2$$

which implies

$$Y_{i_1}'Y_{i_2} = \begin{cases} C_{i_1}^2 & i_1 = i_2 \\ 0 & \text{otherwise} \end{cases}$$

Row-wise PCA

$$\underset{\sim}{X} = LCR'$$

Recall that

$$\hat{\underset{\sim}{X}}_{(j)} = [X_{1(j)},\ldots,X_{n(j)}] = \begin{bmatrix} \cdots & v_{1(j)}' & \cdots \\ \cdots & v_{2(j)}' & \cdots \\ & \vdots & \\ \cdots & v_{p(j)}' & \cdots \end{bmatrix}$$

Just as before,

$$\sum_{i=1}^{p} \|\hat{V}_{i(j)}\|^2 = \sum_{h=1}^{j} c_h^2$$
$$\sum_{i=1}^{p} \|\overset{\perp}{V}_{i(j)}\|^2 = \sum_{h=j+1}^{r} c_h^2$$

$L(R_{(j)})$ is the optimum j-dimensional projection subspace of $\mathbb{R}^2$.

19.2. **Microarray "Prostate Cancer".** Find a linear combination of individual gene expression measurements that are good predictions of overall behaviour.

$$\underbrace{\underset{\sim}{X}}_{p\times n} = [X_1,\ldots,X_n] = LCR' = \sum_{h=1}^{r} c_h l_h r_h'$$

where the rows are genes and the columns are arrays. The first component

$$\underbrace{Y_1'}_{1\times n} = l_1'\underset{\sim}{X} = c_1 r_1'$$

is a reasonable candidate for "informative linear combination".

Homework (C2.3). *(a) For $\underset{\sim}{Y}_{(j)} = L_{(j)}'\underset{\sim}{X}$, $\underset{\sim}{X}$ with row mean 0 ($\sum_{i=1}^{n} X_i = 0$), calculate the $(p+j)\times(p+j)$ matrix*

$$\begin{bmatrix} \underset{\sim}{X} \\ \underset{\sim}{Y}_{(j)} \end{bmatrix} \begin{bmatrix} \underset{\sim}{X}' & \underset{\sim}{Y}_{(j)}' \end{bmatrix}$$

(b) Calculate $\hat{\underset{\sim}{X}}$, the projection of $\underset{\sim}{X}$ row by row into $L_{row}(\underset{\sim}{Y}_{(j)})$
(c) Calculate $\overset{\perp}{\underset{\sim}{X}}\overset{\perp}{\underset{\sim}{X}}'$, where $\overset{\perp}{\underset{\sim}{X}} = \underset{\sim}{X} - \hat{\underset{\sim}{X}}$

Homework (C2.4). *Suppose* $\operatorname{rank}(\underset{\sim}{X}) = p$ *so*

$$\underset{\sim}{X} = L_{p\times p}C_{p\times p}R_{p\times n}' \qquad p \le n$$

Let $M(a,b) \equiv l_a r_b'$, $a = 1,2,\ldots,p$, $b = 1,2\ldots,n$.
(a) Show that $M(a,b)$ are of unit length and mutually orthogonal (in matrix sense)
(b) What is the coordinates of $\underset{\sim}{X}$ in terms of $M(a,b)$?
(c) calculate $\hat{\underset{\sim}{X}}$, the projection of $\underset{\sim}{X}$ row by row into $L_{row}(\underset{\sim}{Y}_{(j)}$.

C.3 Simultaneous Diagonalization and Fisher's LDA.

- Consider again the ratio of quadratic forms

$$Q(g) = \frac{g'Ag}{g'Bg} = \frac{\|g\|_A^2}{\|g\|_B^2}$$

where A and B are $p \times p$ symmetric matrices, $A \geq 0$, $B > 0$. We can rewrite this by defining $\tilde{g} = B^{\frac{1}{2}} g$, or $g = B^{-\frac{1}{2}}\tilde{g}$, where $B^{\frac{1}{2}}$ is a symmetric square root of B. Then

$$Q(b) = \frac{\tilde{g}'(B^{-\frac{1}{2}}AB^{-\frac{1}{2}})\tilde{g}}{\tilde{g}'\tilde{g}} = \frac{\tilde{g}'\tilde{A}\tilde{g}}{\tilde{g}'\tilde{g}}$$

- B is non-singular, $\tilde{A}$ symmetric, $\text{rank}(\tilde{A}) = \text{rank}(A)$.

Simultaneous Diagonalization. Write:

$$\tilde{A} \equiv B^{-\frac{1}{2}}AB^{-\frac{1}{2}} = \Gamma\Lambda\Gamma' \qquad (\star)$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$$

with

$$\lambda_{r+1} = \cdots = \lambda_p = 0$$

$r = \text{rank}(A)$ and $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_p)$, eigenvectors of $\tilde{A}$.

Let

$$\xi = B^{-\frac{1}{2}}\gamma_i$$

Define

$$\Xi = (\xi_1, \ldots, \xi_p) = B^{-\frac{1}{2}}\Gamma$$

Then

$$\Xi' A \Xi = \Gamma' \underbrace{(B^{-\frac{1}{2}}AB^{-\frac{1}{2}})}_{=\tilde{A}=\Gamma\Lambda\Gamma'} = \Lambda$$

and

$$\Xi' B \Xi = \underbrace{\Gamma' B^{-\frac{1}{2}}}_{\Xi'} B \underbrace{B^{-\frac{1}{2}}\Gamma}_{\Xi} = I_p$$

Homework (C3.1)**.** *Verify this*

In other words, $\Xi_{p\times p} = (\xi_1, \ldots, \xi_p)$ diagonalizes A to Λ and B to I_p; the vectors ξ_i are orthonormal in the B metric, and orthogonal in the A metric.

$$\langle \xi_i, \xi_j \rangle_B = \delta_{ij} \qquad \langle \xi_i, \xi_j \rangle_A = \delta_{ij}\lambda_i$$

Metric Eigenvalues. Rewrite the previous results as follows:

$$B\Xi = \Xi'$$
$$A\Xi = \Xi'\Lambda$$

so we have

$$A\Xi = B\Xi\Lambda$$

Comparing the ith column,

$$A\xi_i = \lambda B\xi_i$$

for $1 \leq i \leq p$.

- The values $\lambda_1, \ldots, \lambda_p$ are called the "eigenvalues" of A in the B-metric.
- From the ordinary definition of λ_i, λ_i are the solutions of

$$|\tilde{A} - \lambda I_p| = 0$$

which is equivalent to

$$|B^{-\frac{1}{2}}AB^{-\frac{1}{2}} - \lambda I| = 0$$

which again is equivalent to

$$|A - \lambda B| = 0$$

Homework (C3.2). *For $C_{p\times q}$, show that the non-zero roots of*

$$|\underbrace{CC'}_{p\times p}-\lambda I_p|=0$$

equals the non-zero roots of

$$|\underbrace{C'C}_{q\times q}-\lambda I_q|=0$$

Hint: Use SVD

Homework (C3.3). *If A is non-singular (B is non-singular as before), then λ_i are the solutions of*

$$|B^{-1}-\lambda A^{-1}|=0$$

which implies that λ_i are the eigenvalues of B^{-1} in metric A^{-1}.

20.1. **Restate the Fundamental Lemma.** For $S_{p\times p}$ symmetric, define

$$\overset{\perp}{L}_S(v_1,\dots,v_j)=\{v : v'Sv_h=0, h=1,2,\dots,j\}$$

Return to the ratio

$$Q(b)=\frac{g'Ag}{g'Bg}=\frac{\tilde{g}'\tilde{A}\tilde{g}}{\tilde{g}'\tilde{g}}$$

Then we can restate the fundamental lemma as

(1) $g=\xi_1$ maximizes $Q(g)$
(2) $g=\xi_2$ maximizes $Q(g)$ among all g in $\overset{\perp}{L}_A(\xi_1)$, and all g in $\overset{\perp}{L}_B(\xi_1)$. The maxima are $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$.

Homework (C3.4). *Verify this.*

Theorem 19 (Stationary Theorem). *The vectors $\xi_1,\dots,\xi_p$ are stationary values of*

$$Q(g)=\frac{\|g\|_A^2}{\|g\|_B^2}$$

i.e. they are the solutions of

$$\frac{\partial Q(g)}{\partial g_i}=0$$

for $i=1,2,\dots,p$, ($\nabla Q(g)=0$). If λ_i's are distinct, then $c\xi_i$ are the only stationary values, and $g=\xi_1$ maximizes $Q(g)$ and $g=\xi_p$ minimizes it.

Proof. We can always write

$$\frac{g'Ag}{g'Bg}=\frac{h'\Lambda h}{h'h}$$

where by simultaneous orthogonalization,

$$h=\Gamma' B^{\frac{1}{2}}g=\Gamma'\tilde{g}$$

Now,

$$\frac{h'\Lambda h}{h'h}=\sum\lambda h_i^2$$

on the unit sphere. Now stationary values "obviously" coordinate vectors $h=e_i$, which implies

$$g=B^{-\frac{1}{2}}\Gamma e_i\equiv B^{-\frac{1}{2}}\gamma_i=\xi_i$$

Hint: To find the maximum of the thing above on the unit sphere, we look at $U(h)=h\Lambda h+\lambda(h'h-1)$, the Lagrange form. Our constraint is

$$\frac{\partial U(h)}{\partial h}=0\qquad\frac{\partial U(h)}{\partial\lambda}=0$$

□

Fisher's Linear Discriminant Analysis. Population version: We observe X

$$\begin{cases} X \sim \mathcal{N}(\mu_1, \Sigma) & \text{class 1} \\ X \sim \mathcal{N}(\mu_2, \Sigma) & \text{class 2} \end{cases}$$

Goal: Assume μ_1, μ_2, Σ are given. Given X, we would like to predict the class label Y associated with X to be $Y = 1$ or $Y = 2$.

Sample Version: We have training data sets and test data sets. In the training data sets,

$$\begin{cases} X_1, \ldots, X_{n_1} \sim \mathcal{N}(\mu_1, \Sigma) & n_1 \text{ samples, } Y_1 = \cdots = Y_{n_1} = 1 \\ X_{n_1+1}, \ldots, X_{n_1+n_2} \sim \mathcal{N}(\mu_2, \Sigma) & n_2 \text{ samples, } Y_{n_1+1} = \cdots = Y_{n_1+n_2} = 2 \end{cases}$$

For the test data sets, we have (X, Y) where X is given and Y is unknown.
Goal: predict Y using X and the training set.

One class of methods is called linear discriminant analysis, where we hope to find a one-dimensional linear summary of the form $g'X$

$$Y_1 = g'X_1 \text{ and } Y_2 = g'X_2$$

Fisher's Separation.

$$\frac{\text{Difference of Means}}{\text{SD}}$$

Now

$$g'X_1 \sim \mathcal{N}(g'\mu_1, g'\Sigma g)$$
$$g'X_2 \sim \mathcal{N}(g'\mu_2, g'\Sigma g)$$

so Fisher's separation is

$$\frac{|g'\mu_1 - g'\mu_2|}{\sqrt{g'\Sigma g}}$$

How do we maximize this? Let $\delta = \mu_1 - \mu_2$. Then the Fisher's separation is equal to

$$\frac{|g'\delta|}{\sqrt{g'\Sigma g}} \Rightarrow g \propto \Sigma^{-1}\delta = \Sigma^{-1}(\mu_1 - \mu_2)$$

(By lagrange multipliers, just solve for $(g'\delta)^2 + \lambda(g'\Sigma g - 1)$).

Population Classification Rule: X is a test vector.

$$(\mu_1 - \mu_2)'\Sigma^{-1}X \begin{cases} > t & \text{class 1} \\ < t & \text{class 2} \end{cases}$$

Homework (C3.6)**.** *Verify that* $\max\{Q(g)\} = \delta'\Sigma^{-1}\delta$ *directly from the Fundamental Lemma. Here,*

$$Q(g) = \frac{(g'\delta)^2}{g'\Sigma g} = \frac{g'(\delta\delta')g}{g'\Sigma g}$$

Homework (C3.7)**.** *Suppose we are in a Bayesian situation*

$$X \sim \mathcal{N}(\mu_1, \Sigma)$$

with probability π_0 *or*

$$X \sim \mathcal{N}(\mu_2, \Sigma)$$

with probability π_1*, where* $\pi_0 + \pi_1 = 1$*. The posterior distribution is*

$$\pi_0 f(\mu_1, \Sigma) + \pi_1 f(\mu_2, \Sigma)$$

Therefore,

$$\begin{aligned} \mathbb{P}(X \in \textit{class 1} | X = x) &= \frac{\pi_0 f(\mu_1, \Sigma)}{\pi_0 f(\mu_1, \Sigma) + \pi_1 f(\mu_2, \Sigma)} \\ &= \frac{\pi_0 \frac{1}{|\Sigma|^{p/2}} \exp\left(-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right)}{\frac{1}{|\Sigma|^{p/2}}\left(\pi_0 \exp\left(-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right) + \pi_1 \exp\left(-\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)\right)\right)} \\ &= \frac{\pi_0 e^{\mu_1'\Sigma^{-1}x} e^{-\frac{1}{2}\mu_1'\Sigma^{-1}\mu_1}}{\pi_0 e^{\mu_1\Sigma^{-1}x} e^{-\frac{1}{2}\mu_1'\Sigma^{-1}\mu_1} + \pi_1 e^{\mu_2'x} e^{-\frac{1}{2}\mu_2'\Sigma^{-1}\mu_2}} \\ &= \frac{\pi_0}{\pi_1} e^{\frac{1}{2}(\mu_2'\Sigma^{-1}\mu_2 - \mu_1'\Sigma^{-1}\mu_1} e^{(\mu_1-\mu_2)'\Sigma^{-1}x} \end{aligned}$$

If this is greater than $1/2$*, then we assume it's in class 1. Otherwise, we assume it's in class 2. OR we look at*

$$(\mu_2-\mu_1)'\Sigma(\mu_1+\mu_2)>1$$

$$\exp\left\{(\mu_1-\mu_2)'\Sigma^{-1}\left(x-\frac{\mu_1+\mu_2}{2}\right)\right\}<1$$

We look at

$$\exp\left\{(\mu_1-\mu_2)'\Sigma^{-1}\left(x-\frac{\mu_1+\mu_2}{2}\right)\right\}$$

less then or greater than 1. If we take logs, we can also look at

$$(\mu_1-\mu_2)'\Sigma^{-1}X$$

less than or greater than

$$\log\left(\frac{\pi_1}{\pi_0}\right)+\frac{\mu_2'\Sigma^{-1}\mu_2}{2}-\frac{\mu_1'\Sigma^{-1}\mu_1}{2}$$

21. November 13th, 2013

In the Bayesian model

$$X\sim\mathcal{N}(\mu_1,\Sigma)\quad\text{with prior probability }\pi_1$$
$$X\sim\mathcal{N}(\mu_2,\Sigma)\quad\text{with prior probability }\pi_2$$

where $\pi_1+\pi_2=1$. Letting $\pi_1(x)$ and $\pi_2(x)$ be posterior probabilities,

$$\pi_1(x)=\frac{\pi_1 f_{\mu_1,\Sigma}(x)}{\pi_1 f_{\mu_1,\Sigma}(x)+\pi_2 f_{\mu_2,\Sigma}(x)}$$
$$\pi_2(x)=\frac{\pi_2 f_{\mu_2,\Sigma}(x)}{\pi_1 f_{\mu_1,\Sigma}(x)+\pi_2 f_{\mu_2,\Sigma}(x)}$$
$$\frac{\pi_2(x)}{\pi_1(x)}=\frac{\pi_2 f_{\mu_2,\Sigma}(x)}{\pi_1 f_{\mu_1,\Sigma}(x)}$$

and

$$\lambda(x)=\log\left(\frac{\pi_2(x)}{\pi_1(x)}\right)=\log\left(\frac{\pi_2}{\pi_1}\right)+\beta_0+\beta' x$$

where

$$\beta_0=-\frac{1}{2}(\mu_1'\Sigma^{-1}\mu_1-\mu_2'\Sigma^{-1}\mu_2)$$
$$\beta=\Sigma^{-1}(\mu_2-\mu_1)$$

In the Logistic Model,

$$X_i=\begin{cases}1 & p\\ 0 & 1-p\end{cases}$$

and

$$\log\left(\frac{p}{1-p}\right)=\alpha+\beta' X$$

even when we have samples, logistic model is not efficient, if we use it to estimate β_0, β. We classify

$$\begin{cases}Y=2 & \lambda(x)>0\\ Y=1 & \lambda(x)<0\end{cases}$$

21.1. **Sample Version.** We have $(\mu_1,\mu_2,\Sigma,\pi_1)$ unknown, but we have independent samples (training samples). n_1 of them are in class 1 and n_2 of them are in class 2.

- $\frac{n_2}{n_1}=\frac{\pi_2}{\pi_1}$ is our estimate
- $\hat{\mu}_1=\left(\sum_{i=1}^{n_1}X_{1i}\right)/n_1\sim\mathcal{N}(\mu_1,\Sigma/n_1)$
- $\hat{\mu}_2=\left(\sum_{i=1}^{n_2}X_{2i}\right)/n_1\sim\mathcal{N}(\mu_2,\Sigma/n_2)$
- $\hat{\Sigma}=\frac{1}{n-2}\left[\sum_{i=1}^{n_1}(X_{1i}-\hat{\mu}_1)^2+\sum_{i=1}^{n_2}(X_{2i}-\hat{\mu}_2)^2\right]$

Plug in, obtain $\hat{\beta}_0,\hat{\beta}_j,\log\left(\frac{\hat{\pi}_2}{\hat{\pi}_1}\right)$

$$\hat{\lambda}(x)=\log\left(\frac{\hat{\pi}_2}{\hat{\pi}_1}\right)+\hat{\beta}_0+\hat{\beta}'x$$
$$=\hat{\alpha}+\hat{\beta}'x$$

$CLASSDIAGRAM$

where the point x_0 satisfies

$$\begin{cases} \hat{\alpha} + \hat{\beta} x_0 = 0 \\ x_0 \propto \hat{\beta} \end{cases}$$

21.2. **MKB book.** In Fisher's LDA ($g = 2$), we try to maximize

$$\frac{|(\mu_1 - \mu_2)' w|}{\sqrt{w' \Sigma w}}$$

We have g classes with g means, $\mu_1, \dots, \mu_g$. Define

$$\mu = \frac{\mu_1 + \cdots + \mu_g}{g}$$

In the $g = 2$ case,

$$\begin{aligned} \mu_1 - \mu_2 &\propto \mu_1 - \mu \\ &\propto \mu_2 - \mu \end{aligned}$$

We want to choose w such that

$$\frac{\sum_{i=1}^{g} [w'(\mu_i - \mu)]^2}{w' \Sigma w} \propto \frac{w' B w}{w' \Sigma w}$$

In matrix form, define

$$B = \frac{1}{g} \sum (\mu_i - \mu)(\mu_i - \mu)'$$

which is a $p \times p$ matrix. $W \propto a$, where a is the leading eigenvector of $\Sigma^{-1} B$, a $p \times p$ matrix.

Classify X_i to class k, $k = 1, 2, \dots, g$ if

$$|a' X - a' \bar{X}_k| < |a' X - a' \bar{X}_j|$$

for all $j \neq k$, where a is a test vector

Fisher's LDA ($p \gg n, g = 2$)**.** In a gene microarray, we assume

- Data are centred
- It is equally likely to be in each class ($\pi_1 = \pi_2 = \frac{1}{2}$)

$$\begin{aligned} X_{1i} &\sim \mathcal{N}(-\mu, \Sigma) \\ X_{2i} &\sim \mathcal{N}(\mu, \Sigma) \end{aligned}$$

Call

$$Y_i = \begin{cases} -1 & \text{class 1} \\ +1 & \text{class 2} \end{cases}$$

Then

$$X \sim \mathcal{N}(Y \cdot \mu, \Sigma)$$

Fisher's LDA says

$$L(X) = \underbrace{\mu' \Sigma^{-1}}_{w'} X \begin{cases} < 0 & \text{class 1 } (Y = -1) \\ > 0 & \text{class 2 } (Y = 1) \end{cases}$$

where X is the test feature. The problem here is how do we estimate μ and Σ?

Let us further simplify this problem. Let $\Sigma = I_p$. Assume μ is sparse in the sense that most $\mu_i = 0$. Now, let

$$Z \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i X_i \sim \mathcal{N}(\sqrt{n}\mu, I_p)$$

since if $Y_i = -1$, $X_i \sim \mathcal{N}(-\mu, I_p)$ and if $Y_i = 1$, $X_i \sim \mathcal{N}(\mu, I_p)$.

To estimate μ, the one approach is to take $\sqrt{n}\hat{\mu} = Z$, but then this isn't sparse anymore. Another approach is to take

$$\sqrt{n}\hat{\mu}_i = \begin{cases} Z_i & |Z_i| \geq t \\ 0 & |Z_i| < t \end{cases}$$

or the simplified version

$$\hat{w}_i = \begin{cases} \operatorname{sgn}(Z_i) & |Z_i| \geq t \\ 0 & \text{otherwise} \end{cases}$$

But now our problem is how do we choose t? If t is too large, we might end up with 5 nonzero weights $\hat{w}_i$. If t is too small, we might have 500 nonzero weights.

Procedure for choosing t (Higher Criticism Thresholding).

- For $i = 1, 2, \ldots, p$,
$$\pi_i = \mathbb{P}(|\mathcal{N}(0,1)| > |Z_i|)$$
- Sorting
$$\pi_{(1)} < \pi_{(2)} < \cdots < \pi_{(p)}$$
- Higher Criticism score for classification
$$HC_{p,i} = \sqrt{p}\left(\frac{\frac{i}{p} - \pi_{(i)}}{\sqrt{(\frac{i}{p})(1-\frac{i}{p})}}\right)$$

$$ThresholdingDiagram$$

What is the intuition? For any threshold,
$$\hat{W}_t = \begin{cases} \text{sgn}(Z_i) & |Z_i| \geq t \\ 0 & |Z_i| < t \end{cases}$$

Fisher's Separation is proportional to
$$\frac{|w_t'\mu|}{\sqrt{w_t'w_t}}$$

Classification error: If $\mu_i \sim (1-\varepsilon)\nu_0 + \varepsilon F$,
$$\begin{aligned} C_{err}(t) &= E_{\varepsilon,F}\left(\bar{\Phi}\left(\frac{w_t'\mu}{\sqrt{w_t'w_t}}\right)\right) \\ &\approx \bar{\Phi}\left[E_{\varepsilon,F}\left(\frac{\hat{w}_t'\mu}{\sqrt{\hat{w}_t'\hat{w}_t}}\right)\right] \\ &\approx \bar{\Phi}\left[\frac{E_{\varepsilon,F}(\hat{w}_t\mu)}{\sqrt{E_{\varepsilon,F}(\hat{w}_t'\hat{w}_t)}}\right] \\ &\approx \Phi(\widetilde{Sep}(t)) \end{aligned}$$

where $\bar{\Phi} = 1 - \Phi$, the survival function of the standard normal. ε_p is small and F_p at "right"/"most interesting" scale.

Example 12. *$\varepsilon_p = p^{-\vartheta}$, ($p\varepsilon_p = p^{1-\vartheta}$) where $0 < \vartheta < 1$. Then*
$$F_p = V_{\tau_p}$$
which is a point mass at τ_p, where $\tau_p = \sqrt{2r\log p}$.

Back to the diagram, we have $HC_{p,i}$ converges to another curve called $\widetilde{HC}(\bar{\Phi}(t))$ as $p \to \infty$.
In a second case, Σ is unknown, but $\Omega = \Sigma^{-1}$ is sparse and each row has very few nonzeros.

- Estimate Ω
- Even when Ω is given or can be estimated very well

Recall $Z \sim \mathcal{N}(\sqrt{n}\mu, \Sigma)$, where
$$Z = \sqrt{n}\mu + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \Sigma)$$
Let $\tilde{Z} = \Sigma^{-1/2}Z$, so
$$\tilde{Z} = \Sigma^{-1/2}\sqrt{n}\mu + \tilde{\varepsilon}$$
where
$$\tilde{\varepsilon} = \Sigma^{-1/2}\varepsilon \sim \mathcal{N}(0, I_p)$$
$\min \|Y - X\beta\|^2$, up to $\|\beta\|_1 < s$and $\min \|\beta\|_1$ up to $Y = X\beta$.

Marginal Regression.
$$\Sigma^{-1/2}\tilde{Z} = \Sigma^{-1}(\sqrt{n}\mu) + \mathcal{N}(0, \Sigma^{-1})$$
Then do thresholding:
$$\hat{w}_t = \begin{cases} \text{sgn}((\Omega Z)_i) & |(\Omega Z)_i| \geq t \\ 0 & \text{otherwise} \end{cases}$$

22. November 18th, 2013

Friday December 6th, possible makeup class. Final will be released probably December 8th and we will get until Friday, the 13th to complete it.

- $X_i \sim \mathcal{N}(Y_i\mu, I_p)$, $Y_i \pm 1$ equally likely
  $X \sim \mathcal{N}(Y_1\mu, I_p)$ Y unknown, $Y \in \{-1, 1\}$. X given.
- Procedure:

$$L(X) = W_t'X \begin{cases} > 0 & \hat{Y} = 1 \\ < 0 & \hat{Y} = -1 \end{cases}$$

We have

$$W_t' = \begin{cases} Z_i & |Z_i| \geq t \\ 0 & \text{otherwise} \end{cases}$$

where $Z = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i X_i$. The value t is determined by Higher Criticism.

Now if $\Sigma = I_p$, Fisher's LDA says

$$\begin{aligned} L(X) &= \mu'\Sigma^{-1}X \\ &= \mu'X \end{aligned}$$

Now, if μ is unknown,

- We estimate μ by $Z/\sqrt{n}$, which gives too many nonzeros.
- Fix the problem by

$$\hat{\mu} = W_t$$

If $\Sigma \neq I_p$, Fisher's LDA would have

$$L(X) = \mu'\Sigma^{-1}X$$

There are two problems

- How do we estimate Σ^{-1}? (We can do glasso possibly.)
- How do we estimate μ?

μ is sparse and $\Omega = \Sigma^{-1}$ is also sparse.

Problem: Assume $\Omega \equiv \Sigma^{-1}$ is give and sparse. How do we estimate μ, with a tuning-parameter free algorithm?
Approach: $Z \sim \mathcal{N}(\sqrt{n}\mu, \Sigma)$. Then we can rewrite this as

$$Z = \sqrt{n}\mu + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \varepsilon)$$

or also

$$\Sigma^{-\frac{1}{2}}Z = \sqrt{n}\Sigma^{-\frac{1}{2}}\mu + \underbrace{\Sigma^{-\frac{1}{2}}\varepsilon}_{\sim\mathcal{N}(0,I_p)}$$

$Y = X(\sqrt{n}\mu)t$

Marginal Regression: If $Y = X\beta + \varepsilon$, then

$$x'y = \begin{bmatrix} (x_1, y) \\ (x_2, y) \\ \vdots \\ (x_p, y) \end{bmatrix}$$

where $x = [x_1, x_2, \ldots, x_p]$. The pretend the true model is $y = \beta_i x_i + \varepsilon$.

For $t > 0$, we estimate

$$\hat{\beta}_i = \begin{cases} (x_i, y) & \text{if } |(x_i, y)| > t \\ 0 & \text{otherwise} \end{cases}$$

This tells us $\beta = \sqrt{n}\mu$ and $X = \Sigma^{-\frac{1}{2}}$, which imply

$$\sqrt{n}\hat{\mu}_i = \begin{cases} (\Sigma^{-1}Z)_i & |(\Sigma^{-1}Z)_i| \geq t \\ 0 & \text{otherwise} \end{cases}$$

How do we determine t, the threshold value? We can use Higher Criticism. Assume Ω is standardized such that the diagonals are 1.

- $\pi_i = \mathbb{P}\big(|\mathcal{N}(0,1)| \geq |(\Sigma^{-1}Z)_i|\big)$
- $\pi_{(1)} < \pi_{(2)} < \vdots < \pi_{(p)}$

- Let $\hat{i}$ maximize

$$\sqrt{p}\frac{\frac{i}{n}-\pi_{(i)}}{\sqrt{\frac{i}{n}(1-\frac{i}{n})}}$$

- Same as before.

C.4 Critical Angles and Canonical Correlations.

- Simultaneous diagonalization theory gives an elegant statement of the geometric relationship between two linear spaces in $\mathbb{R}^n$
- Suppose

$$\underset{\sim}{X}_{p\times n}=\begin{bmatrix}\cdots & v_1' & \cdots\\ \cdots & v_2' & \cdots\\ & \vdots & \\ \cdots & v_p' & \cdots\end{bmatrix}$$

  is full rank (p) with $p \le n$.
- Define $L_B = L_{row}(\underset{\sim}{X})$ as a p-dimensional subspace of $\mathbb{R}^n$.
- Suppose we have another space

$$L_A = \text{a } q\text{-dimensional subspace of } \mathbb{R}^n$$

$$ANGLEDIAGRAM$$

- Goal: Find $v \in L_B$ such that v and $P_A v$ has an angle as small as possible. Call this v_1.
- Find $v_2 \in L_B$ such that $(v_2, v_1) = 0$ and v_2 and $P_A v_2$ have an angle as small as possible.
- The row by row projection of $\underset{\sim}{X}$ into L_A is

$$\underset{\sim}{\hat{X}} = \underset{\sim}{X}P_A = \begin{bmatrix}\cdots & \hat{v}_1' & \cdots\\ \cdots & \hat{v}_2' & \cdots\\ & \vdots & \\ \cdots & \hat{v}_p' & \cdots\end{bmatrix}$$

  A vector

$$\underbrace{v'}_{1\times n} = g' \underbrace{\underset{\sim}{X}}_{p\times n} \in L_{row}(\underset{\sim}{X}) = L_B$$

  in L_B projects into $\hat{v}' = g'\underset{\sim}{X}P_a = g'\underset{\sim}{\hat{X}}$
- The $\cos^2$ of the "θ" between v and $\hat{v}$ is then

$$\begin{aligned}\cos^2\theta &= \frac{\|\hat{v}\|^2}{\|v\|^2}\\ &= \frac{g'\underset{\sim}{\hat{X}}\underset{\sim}{\hat{X}}'g}{g'\underset{\sim}{X}\underset{\sim}{X}'g}\\ &= \frac{g'\underset{\sim}{X}P_A^2\underset{\sim}{X}'g}{g'\underset{\sim}{X}\underset{\sim}{X}'g}\\ &\equiv \frac{g'Ag}{g'Bg}\\ &\equiv Q(g)\end{aligned}$$

A and B are symmetric, positive semi-definite matrices. note that

$$C = B - A = \underset{\sim}{X}(I_n - P_A)\underset{\sim}{X}' \ge 0$$

Moreover, $B^{-\frac{1}{2}}CB^{-\frac{1}{2}} = I - B^{-\frac{1}{2}}AB^{-\frac{1}{2}} \ge 0$. Therefore, the roots of $\tilde{A} = B^{-\frac{1}{2}}AB^{-\frac{1}{2}}$ satisfy

$$1 \ge \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p \ge 0$$

Homework (C4.1)**.** *Verify this.*

Applying the simultaneous diagonalization theory, let

$$\tilde{A} = B^{-\frac{1}{2}}AB^{-\frac{1}{2}} = \Gamma\Lambda\Gamma'$$

and

$$\Xi = B^{-\frac{1}{2}}\Gamma = [\xi_1, \ldots, \xi_p]$$

We have

$$\Gamma = (\gamma_1, \ldots, \gamma_p)$$
$$\xi_h = B^{-\frac{1}{2}}\gamma_h \quad h = 1, 2, \ldots, p$$

Theorem 20.

- $g = \xi_1$, *or* $v_1 = \xi_1' \underset{\sim}{X}$ *maximizes* $\cos^2(\theta(g))$ *with* $\cos\theta_1 = \sqrt{\lambda_1}$
- $g = \xi_2$, *or* $v_2 = \xi_2' \underset{\sim}{X}$ *maximizes* $\cos^2(\theta(g))$, *subject to* $g'A\xi_1 = 0$ *or* $g'B\xi_1 = 0$.

Orthogonal Pairs.

$$\underbrace{\underset{\sim}{Y}}_{p\times n} = \Xi'\underset{\sim}{X} = \begin{bmatrix} \cdots & y_1' & \cdots \\ \cdots & y_2' & \cdots \\ & \vdots & \\ \cdots & y_p' & \cdots \end{bmatrix}$$

so that $y_n' = \xi_n' \underset{\sim}{X}$ and

$$\underbrace{\underset{\sim}{\hat{Y}}}_{p\times n} = \Xi' \underbrace{\underset{\sim}{\hat{X}}}_{=\underset{\sim}{X}P_A} = \Xi'\underset{\sim}{X}P_A = \underset{\sim}{Y}P_A$$

As before, we get orthogonality in both metrics.

$$\underset{\sim}{Y}\underset{\sim}{Y}' = \Xi'\underset{\sim}{X}\underset{\sim}{X}'\Xi = \Xi' B \Xi = I_p$$

and

$$\underset{\sim}{\hat{Y}}\underset{\sim}{\hat{Y}}' = \Xi'\underset{\sim}{X}P_A\underset{\sim}{X}'\Xi = \Xi' A \Xi = \Gamma'\tilde{A}\Gamma = \Lambda$$

Thus we can write

$$\begin{bmatrix} \underset{\sim}{Y} \\ \underset{\sim}{\hat{Y}} \end{bmatrix} \begin{bmatrix} \underset{\sim}{Y}' & \underset{\sim}{\hat{Y}}' \end{bmatrix} = \begin{bmatrix} I_p & \Lambda \\ \Lambda & \Lambda \end{bmatrix}$$

You have p pairs of vectors

$$(y_h, \hat{y}_h), \quad h = 1, 2, \ldots, p$$

- Different pairs are orthogonal pairs
- Within the pair, the second one is the orthogonal projection of the first one.

Homework (C4.2)**.** *Show that the off diagonal "*Λ*" is correct.*

Summary: We can restate things geometrically as follows. There are p pairs, $(y_h, \hat{y}_h)$ where $y_h \in L_B = L_{row}(\underset{\sim}{X})$ and $\hat{y}_h = P_A y_h$ such that

(i) The t_h are mutually orthonormal. ($\|y_h\| = 1$)
(ii) $\hat{y}_h$ are mutually orthogonal ($\|\hat{y}_h\|^2 = \lambda_h$)
(iii) All $2p$ vectors are mutually orthogonal outside the pairs

$$(y_h, y_{h'}) = (y_h, \hat{y}_{h'}) = (\hat{y}_h, \hat{y}_{h'}) = 0$$

if $h \neq h'$.

(iv) The smallest possible angle is between y_1 and $\hat{y}_1$ ($g'\underset{\sim}{X}$ and $g'\underset{\sim}{\hat{X}}$), with $\cos^2(\theta_1) = \lambda_1$. The smallest possible angle between $g'\underset{\sim}{X}$ and $g'\underset{\sim}{\hat{X}}$ subject to to "$\perp$" to y_1 is y_2 and $\hat{y}_2$. The smallest angle between $g'\underset{\sim}{X}$ and $g'\underset{\sim}{\hat{X}}$ is given by y_p and $\hat{y}_p$, $\cos^2(\theta_p)$.

23. November 25th, 2013

Let $L_b = L_{row}(\underset{\sim}{X})$ where

$$\underset{\sim}{X} = \begin{pmatrix} \cdots & v_1' & \cdots \\ \cdots & v_2' & \cdots \\ & \vdots & \\ \cdots & v_p' & \cdots \end{pmatrix}$$

and A is another space.
How do we define angles between spaces?

As before, we stick with the same notation

$$\tilde{A} = B^{-\frac{1}{2}} A B^{-\frac{1}{2}}$$
$$B = \underset{\sim}{X}\underset{\sim}{X}'$$
$$A = \underset{\sim}{X}P_A\underset{\sim}{X}'$$

and

$$\tilde{A} = \Gamma\Lambda\Gamma' \qquad \Xi = B^{-\frac{1}{2}}\Gamma \qquad \underbrace{\underset{\sim}{Y}}_{p\times n} = \underbrace{\Xi'}_{p\times p}\underbrace{\underset{\sim}{X}}_{p\times n}$$

Last time we had p pairs, where each pair is 2 dimensional. $(y_h, \hat{y}_h)$ such that

- $(y_h, \hat{y}_h) = \cos\theta_h$
- Different pairs are orthogonal
- θ_1 is the smallest and θ_p is the largest (they are ordered) These are the "critical angles".

Remark: $\theta_1, \dots, \theta_p$ do not depend on the basis we choose in L_B (the matrix $\underset{\sim}{X}$). Also remark that

$$y_h \xrightarrow{P_A} \hat{y}_h \xrightarrow{P_B} \cos^2\theta_h y_h$$

For simplicity, let us say $\lambda_h > 0$. Then

$$\|\hat{y}_h\|^2 = \lambda_h > 0$$

Lemma. *The projection of $\hat{Y} = \underset{\sim}{Y}P_A$ back to L_B is*

$$\hat{\underset{\sim}{Y}} = \Lambda\underset{\sim}{Y} = \begin{pmatrix} \cdots & \cos^2\theta_1 Y_1 & \cdots \\ \cdots & \cos^2\theta_2 Y_2 & \cdots \\ & \vdots & \\ \cdots & \cos^2\theta_p Y_p & \cdots \end{pmatrix}$$

Proof. Since the rows of $\underset{\sim}{Y}$ are orthogonal,

$$\underbrace{\underset{\sim}{Y}}_{p\times n}\underbrace{\underset{\sim}{Y}'}_{n\times p} = I_p$$

we can take the projection operator into L_B to be

$$\underbrace{P_B}_{n\times n} = \underbrace{\underset{\sim}{Y}'}_{n\times p}\underbrace{\underset{\sim}{Y}}_{p\times n}$$

Then

$$\begin{aligned} \hat{\underset{\sim}{Y}} &= \hat{\underset{\sim}{Y}}P_B \\ &= \underset{\sim}{Y}P_A P_B \\ &= \underbrace{\Xi'\underset{\sim}{X}}_{\underset{\sim}{Y}} P_A \underset{\sim}{Y}'\underset{\sim}{Y} \\ &= \Xi'\underset{\sim}{X}P_A\underset{\sim}{X}\Xi\underset{\sim}{Y} \\ &= (\Xi' A \Xi)\underset{\sim}{Y} \\ &= \Lambda\underset{\sim}{Y} \end{aligned}$$

In conclusion,

$$\hat{\underset{\sim}{Y}} = \Lambda\underset{\sim}{Y}$$

Recall that $\lambda_h = \cos^2\theta_h$. □

Defining this symmetrically, we are going to normalize $\hat{Y}_h$. (Remember that $\|y_h\| = 1$ and $\|\hat{y}_h\| = \lambda_h$). Let

$$y_{h,A} = \frac{\hat{Y}_h}{\sqrt{\lambda_h}}$$

so that $\|y_{h,A}\| = 1$ and

$$y_{h,B} = y_h$$

Now, $\underset{\sim}{Y}_A$ which is $p \times n$ is equal to

$$\underset{\sim}{Y}_A = \left(\cdots y'_{h,A} \cdots\right)_{h=1,2,\dots,p}$$

and

$$\underset{\sim}{Y}_B = \left(\cdots y'_{h,B} \cdots\right)_{h=1,2,\dots,p}$$

We have

$$\begin{bmatrix} \underset{\sim}{Y}_B \\ \underset{\sim}{Y}_A \end{bmatrix}\begin{bmatrix} \underset{\sim}{Y}'_B & \underset{\sim}{Y}'_A \end{bmatrix} = \begin{bmatrix} I_p & \Lambda^{\frac{1}{2}} \\ \Lambda^{\frac{1}{2}} & I_p \end{bmatrix}$$

Homework (C4.3)**.** *Verify this*

In summary: Given $\underset{\sim}{X}$ we have $B = \underset{\sim}{X}\underset{\sim}{X}'$ and $A = \underset{\sim}{X}P_A\underset{\sim}{X}' = \Gamma\Lambda\Gamma'$, which implies $\Xi = B^{-\frac{1}{2}}\Gamma$ and $\underset{\sim}{Y} = \Xi'\underset{\sim}{X}$

(i) The spaces $L_h = L(y_{h,A}, y_{h,B})$ are mutually orthogonal
(ii) $L_A = (y_{1,A}, y_{2,A}, \ldots, y_{p,A})$ and $L_B = (y_{1,B}, y_{2,B}, \ldots, y_{p,B})$
(iii) The angles between y_h and $\hat{y}_h$ are $\cos^2\theta_h = \lambda_h$
(iv) $\{y_{h,A}\}$ and $\{y_{h,B}\}$ are orthonormal
(v) The θ_i's are critical angles. θ_1 is the smallest. θ_2 is the second smallest under constraints... and θ_p is the largest one

Homework (C4.4)**.** *What happens if for some h, $\lambda_h = 0$? If $p \neq q$? Give a geometric description. (Jin said that this doesn't require much proof)*

Homework (C4.5)**.** *Suppose instead of $\underset{\sim}{X}$, we begin with $\tilde{\underset{\sim}{X}} = M\underset{\sim}{X}$ where $\tilde{\underset{\sim}{X}}$ is $p \times n$, M is $p \times p$ and $\underset{\sim}{X}$ is $p \times n$, so that $L_{row}(\tilde{\underset{\sim}{X}}) = L_{row}(\underset{\sim}{X})$. Show that the pairs of $(\tilde{y}_{h,A}, \tilde{y}_{h,B})$ are the same as $(y_{h,A}, y_{h,B})$. In other words, that it does not matter how L_A is described. (In particular, take $M = \Xi'$, $\tilde{\underset{\sim}{X}} = \underset{\sim}{Y}$). Hint: consider $A - \lambda B$ as before*

23.1. **Projection Ratios and Cauchy Projection Formula.** Suppose L_B and L_G are subspaces of $\mathbb{R}^n$, with dimension $p \leq q$. Let "c" be a figure in L_B with point by point projection

$$\hat{c} = \{\hat{v}, v \in c\}$$

into L_G.

$$CDIAGRAM$$

Lemma. *The p-dimensional volumes are related by*

$$\text{Vol}(\hat{C}) = \text{Vol}(C)\prod_{h=1}^{p}\cos(\theta_h)$$

where $\theta_1, \ldots, \theta_p$ are the critical angles between L_B and L_G.

Sketch of the proof. Because the mapping $v \to \hat{v} = vP_A$ is linear, then if $c = \bigcup_k c_k$, where c_k are small cubes.

$$\text{Vol}(\hat{C}_k) = \text{Vol}(C_k)\prod_{h=1}^{p}\cos(\theta_h)$$

□

Homework (C5.1)**.** *Verify*

Let

$$\underbrace{\underset{\sim}{Z}}_{p\times n} = \begin{bmatrix} \cdots & z_1' & \cdots \\ \cdots & z_2' & \cdots \\ & \vdots & \\ \cdots & z_p' & \cdots \end{bmatrix}$$

be an orthonormal set of L_B.

$$\underset{\sim}{Z}\underset{\sim}{Z} = I_P \quad \text{and} \quad L_{row}(\underset{\sim}{Z}) = L_B$$

Then c, the cube whose lower corner is defined by $\underset{\sim}{Z}$ has $\text{Vol}^2(c) = 1$.The projection $\hat{c}$, defined by

$$\hat{\underset{\sim}{Z}} = \begin{bmatrix} \cdots & \hat{z}_1 & \cdots \\ \cdots & \hat{z}_2 & \cdots \\ & \vdots & \\ \cdots & \hat{z}_p & \cdots \end{bmatrix} = \underset{\sim}{Z}P_A$$

has

$$\begin{aligned} \text{Vol}^2(\hat{c}) &= |\hat{\underset{\sim}{Z}}\underset{\sim}{Z}'| \\ &= |\Gamma\Lambda\Gamma'| \\ &= |\Lambda| \\ &= \prod_{h=1}^{p}\lambda_h \end{aligned}$$

Here the notation differs. Before, we had $L_B = L_{row}(\underset{\sim}{X})$ and L_A is any space of p dimensions. Now, $L_B = L_{row}(\underset{\sim}{Z})$. Then $A = \underset{\sim}{Z}P_A\underset{\sim}{Z}' = \Gamma\Lambda\Gamma'$.

Homework (C5.2)**.** *Prove this.*

Then $|\hat{\underset{\sim}{Z}}\hat{\underset{\sim}{Z}}'| = \prod_{h=1}^{p} \lambda_h$, where λ_h are eigenvalues of $\hat{\underset{\sim}{Z}}\hat{\underset{\sim}{Z}}'$.

$$0 = |\hat{\underset{\sim}{Z}}\hat{\underset{\sim}{Z}}' - \lambda I_p| = |\hat{\underset{\sim}{Z}}\hat{\underset{\sim}{Z}} - \lambda \underset{\sim}{Z}\underset{\sim}{Z}'|$$

for $\lambda = \lambda_h$. Define

$$R = \prod_{h=1}^{p} \cos(\theta_h) = \prod_{h=1}^{p} \lambda_h^{\frac{1}{2}}$$

which is the projection ratio from L_B to L_G.

Critical angles

- Do not depend on the basis we choose in L_B.
- Do not matter whether the projection is L_A to L_B or L_B to L_A.

23.2. **Distribution Theory for Projection Ratio R.** Related to the Beta random variable. $Z \sim \text{Beta}(\alpha, \beta)$ has density

$$f_Z(z) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

for $0 < x < 1$ with moments

$$\mathbb{E}[Z^m] = \frac{\Gamma(\alpha+m)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+m)}$$

Theorem 21. *Suppose we have L_a and L_b where L_a has dimension q and is fixed and L_b is of dimension p and is random, uniformly distributed in directions, i.e. invariably distributed under orthogonal transformation in $\mathbb{R}^n$.*

Then R^2 is distributed as the product of independent Beta random variables

$$R^2 = \prod_{h=1}^{p} \text{Beta}\left(\frac{q-h+1}{2}, \frac{n-q}{2}\right)$$

Proof. Without loss of generality, we can take $L_a = L(e_1, e_2, \ldots, e_q)$ and $L_b = L_{row}(\underset{\sim}{X})$ and $\underset{\sim}{X} \sim \mathcal{N}_{p\times n}(\mathbf{0}, I_p \bigotimes I_n)$.

Then $\hat{X} = (X_{(q)}, 0)$ where $X_{(q)}$ is $p \times q$ and the 0 part of the matrix is $p \times (n-q)$. As in the proof of the lemma earlier,

$$R^2 = \frac{|\hat{\underset{\sim}{X}}\hat{\underset{\sim}{X}}'|}{|\underset{\sim}{X}\underset{\sim}{X}'|} = \frac{|X_{(q)}X_{(q)}'|}{|\underset{\sim}{X}\underset{\sim}{X}'|}$$

$X_{(q)}$ is made of the first q columns of $\underset{\sim}{X}$. Now, write $\underset{\sim}{X}$ in triangular-orthogonal form.

$$\underbrace{\underset{\sim}{X}}_{p\times n} = \underbrace{T}_{p\times p}\underbrace{\underset{\sim}{W}}_{p\times n}$$

where $\underset{\sim}{W}\underset{\sim}{W}' = I_p$. Recall, $\underset{\sim}{W} \sim$"Uniform" and independent of T. We have

$$\underbrace{\underset{\sim}{X}_{(q)}}_{p\times q} = \underbrace{T}_{p\times p}\underbrace{\underset{\sim}{Y}_{(q)}}_{p\times q}$$

where $\underset{\sim}{Y}_{(q)}$ is the first q columns of W and

$$\begin{aligned} R^2 &= \frac{|T\underset{\sim}{Y}_{(q)}\underset{\sim}{Y}_{(q)}'T'|}{|T\underset{\sim}{W}\underset{\sim}{W}'T|} \\ &= \frac{|T|^2|\underset{\sim}{Y}_{(q)}\underset{\sim}{Y}_{(q)}'|}{|T|^2|\underset{\sim}{W}\underset{\sim}{W}'|} \\ &= |\underset{\sim}{Y}_{(q)}\underset{\sim}{Y}_{(q)}'| \end{aligned}$$

independent of T, for $W \perp\!\!\!\perp T$.

Now,

(a) $|\underset{\sim}{X}\underset{\sim}{X}'| = |T\underset{\sim}{W}\underset{\sim}{W}'T| = |T^2|$

(b) $\underbrace{|\underset{\sim}{X}\underset{\sim}{X}|}_{T^2} R^2 = |\underset{\sim}{X}_{(q)}\underset{\sim}{X}_{(q)}'|$

where T and R are independent, so

$$\mathbb{E}\left[|\underset{\sim}{X}\underset{\sim}{X}|^{\frac{m}{2}}\right]\mathbb{E}\left[(R^2)^{\frac{m}{2}}\right] = \mathbb{E}\left[|\underset{\sim}{X}_{(q)}\underset{\sim}{X}_{(q)}'|^{\frac{m}{2}}\right]$$

Now, $\underset{\sim}{X}\underset{\sim}{X}' \sim W(I_p; p, n)$, $\underset{\sim}{X}_{(q)}\underset{\sim}{X}_{(q)}' \sim W(I_p, p, q)$. □

Homework (C5.3)**.** *Use this to show for $m > 0$*

(a)

$$\mathbb{E}[R^m] = \prod_{h=1}^{p} \frac{\Gamma(\frac{q+m-h+1}{2})\Gamma(\frac{n-h+1}{2})}{\Gamma(\frac{q-h+1}{2})\Gamma(\frac{n+m-h+1}{2})}$$

(b) Why does this give the theorem?

24. December 2nd, 2013

In the population version, we have $\mu_1, \mu_2, \mu_3, \Sigma$ are given. We have (X, Y) where X is given and Y is unknown. Say we have π_1, π_2, π_3. And $c(i|j)$ is the cost of mistakenly classifying a label j as a label i. This is nonzero for $i \neq j$. $c(i|i) = 0$.

$$\begin{aligned}\mathbb{E}\Big[\text{Cost}(Y, \hat{Y})\Big] &= \sum_{i=1}^{3} \pi_i \mathbb{E}\Big[\text{Cost}(Y, \hat{Y})|Y = i\Big] \\ &= \pi_1 \left[c(2|1)\mathbb{P}\Big(\hat{Y} = 2|Y = 1\Big) + c(3|1)\mathbb{P}\Big(\hat{Y} = 3|Y = 1\Big)\right] \\ &\quad + \cdots \\ &\quad + \cdots\end{aligned}$$

Homework (1)**.** *Suppose we have weights $c(i, j)$, which is constant whenever $i \neq j$ and equal to 0 when $i = j$. Show this gives Bayesian classification rule.*

Homework (2)**.** *Extend it to the unequal weight case.*

Definition 13.

$$R = \prod_{h=1}^{p} \cos\theta_h = \prod_{h=1}^{p} \lambda_h^{\frac{1}{2}}$$

is the projection ratio from L_B to L_A where $\theta_1, \theta_2, \ldots, \theta_p$ are the critical angles.

24.1. Distribution Theory.

Theorem 22. *Suppose L_a which is fixed with dimension $q \geq 1$ and L_B with dimension p are random, uniformly distributed in direction. Then R^2 is distributed as the product of independent Beta's and*

$$R^2 \sim \prod_{h=1}^{p} \text{Beta}\left(\frac{q-h+1}{2}, \frac{n-q}{2}\right)$$

Proof Sketch:

$$\begin{aligned}L_a &= L(e_1, \ldots, e_q) \\ L_B &= L_{row}(\underset{\sim}{X}), \underset{\sim}{X} \sim \mathcal{N}_{p\times n}(\mathbf{0}, I_p \bigotimes I_n)\end{aligned}$$

This way,

$$\hat{\underset{\sim}{X}} = (\underset{\sim}{X}_{(q)}, \mathbf{0})$$

where $\underset{\sim}{X}_{(q)}$ is $p \times (n - q)$. Now, by triangular decomposition,

$$\begin{aligned}\underset{\sim}{X} &= TW \\ \underset{\sim}{X}_{(q)} &= TY_{(q)}\end{aligned}$$

where $Y_{(q)}$ is the first q columns of W

- $R^2 = |Y_{(q)}Y'_{(q)}|$, independent of T^2
- $|\underset{\sim}{X}\underset{\sim}{X}'| = T^2$
- $R^2|\underset{\sim}{X}\underset{\sim}{X}'| = |\underset{\sim}{X}_{(q)}\underset{\sim}{X}'_{(q)}|$.

□

In HWC5.3,

$$\mathbb{E}[R^m] = \prod_{h=1}^{p} \frac{\Gamma(\frac{q+m-h+1}{2})\Gamma(\frac{n-h+1}{2})}{\Gamma(\frac{q-h+1}{2})\Gamma(\frac{n+m-h+1}{2})}$$

when $m = 1$,

$$\mathbb{E}[R] = \prod_{h=1}^{p} \frac{\Gamma(\frac{q-h+2}{2})\Gamma(\frac{n-h+1}{2})}{\Gamma(\frac{q-h+1}{2})\Gamma(\frac{n-h+2}{2})}$$

Definition 14. *Call* $\mathbb{E}[R]$ *by* $\gamma_n(p,q)$*. Moreover, when* $p=q=n-1$*,*

$$\gamma_n(p,q)=\gamma_n(n-1,n-1)=\prod_{h=1}^{n-1}\frac{\left[\Gamma(\frac{n-h+1}{2})\right]^2}{\Gamma(\frac{n-h}{2})\Gamma(\frac{n-h+2}{2})}$$

and since $\Gamma(1+x)=x\Gamma(x)$*, this further equals*

$$\prod_{h=1}^{n-1}\left(\frac{n-h}{2}\right)\left[\frac{\Gamma(\frac{n-h+1}{2})}{\Gamma(\frac{n-h}{2})}\right]^2$$

Now, suppose c is a fixed set in L_a. Let $\hat{c}_B$ be its projection into L_B with p-volume $\hat{V}_B$. Since $\hat{V}_B=VR_B$, the average projected volume is then

$$\begin{aligned}\bar{V}&=\int VR_B\mathrm{d}U(L_B)\\&=V\mathbb{E}[R_B]\\&=\gamma_n(p,q)V\end{aligned}$$

where L_B is chosen "uniformly" in direction, which implies

$$V=\frac{\bar{V}}{\gamma_n(p,q)}$$

24.2. **Cauchy's Projection Formula.** A given convex body in $\mathbb{R}^n$ has $(n-1)$-dimensional "surface area" V. Its projection into a random $(n-1)$-dimensional subspace has $(n-1)$-dimensional volume $\hat{V}_a$. If $\bar{V}$ is the average of $\hat{V}_a$, then

$$V=\frac{2\bar{V}}{\gamma_n(n-1,n-1)}$$

Homework (C5.4)**.** *Prove this for the case of* $n=2$*, assuming* c *is a convex polytope. Where does "2" come from?*

24.3. **Clustering.** Clustering is very similar to classification. In classification, you have a training set

$$(X_i,Y_i),Y_i\in\{1,2,\dots,g\}$$

which are given and a test set (X,Y) where X is given and Y is unknown. In theory g, the number of classifications, is usually unknown, but in practice it is unknown.

In clustering, we have $(X_i,Y_i),i=1,2,\dots,n$ where $X_i\in\mathbb{R}^p$, $Y_i\in\{1,2,\dots,g\}$.

24.4. **Normal Theory.** $X_i\sim\mathcal{N}(\mu_k,\Sigma)$, if i is in the kth class. We call l a $n\times 1$ label vector if $l_i\in\{1,2,\dots,g\}$.

$$\begin{aligned}c_1&=\{1\le i\le n,l_i=1\}\\&\vdots\\c_g&=\{1\le i\le n,l_i=g\}\end{aligned}$$

Now, give an l, we have a likelihood

$$\prod_{k=1}^{g}\prod_{i\in c_k}f(x_i|\mu_k,\Sigma)$$

and the log-likelihood is

$$c-\frac{n}{2}\log|\Sigma|-\frac{1}{2}\sum_{k=1}^{g}\left[\sum_{i\in c_k}(X_i-\mu_k)'\Sigma^{-1}(X_i-\mu_k)\right]$$

where c is a constant.

- $\hat{\mu}_k=\hat{\mu}_k(l)$
- $\hat{\Sigma}$ is the pooled covariance matrix

$$\frac{1}{n-g}\sum_{k=1}^{g}\sum_{i\in c_k}(X_i-\hat{\mu}_k)(X_i-\hat{\mu}_k)'$$

- Plugging in the log-likelihood is then the log-likelihood with $(\hat{\mu},\hat{\Sigma})$ plugged in.
- This can be done if you feed in an l.
- The optimal l is then the one that optimizes the plug-in log-likelihood.

$$\operatorname{tr}\left(\hat{\Sigma}^{-1}\sum_{k=1}^{g}\sum_{i\in c_k}(X_i-\mu_k)(X_i-\mu_k)'\right)$$

$n\to\infty$, p fixed, $n\ll p$.

- k means
- Assume features are uncorrelated/normalized so we don't have to worry about Σ
- Given an l, calculate μ_k as before
- Pick the l to optimize

$$\sum_{k=1}^{g}\sum_{i\in c_k}(X_i-\hat{\mu}_k)^2$$

Decompose

$$\underset{\sim}{X}_{n\times p}=\text{mean}+\text{noise}$$

where the mean is of the form

$$\begin{bmatrix}\mu_1'\\ \mu_3'\\ \mu_1'\\ \mu_2'\\ \vdots\end{bmatrix}$$

which is an $n\times p$ matrix. There are g different kinds of rows.

25. December 4th, 2013

Clustering: very similar to classification. The labels are known in the training set and unknown in the test set?

25.1. **Probability Theory (NP hard).** Normal Theory:

$$X_i\sim\mathcal{N}(\mu_k,\Sigma_k)$$

for $1\le k\le g$. Then, give any label vector γ, which is $m\times 1$, $\gamma_i\in\{1,2,\dots,g\}$. The log-likelihood is

$$C-\frac{1}{2}\sum_{k=1}^{g}\sum_{i\in c_k}(X_i-\mu_k)'\Sigma_k^{-1}(X_i-\mu_k)-\frac{1}{2}\sum_{k=1}^{g}|c_k|\log|\Sigma_k|$$

If you plug this in with $\hat{\mu}_k(\gamma)=\bar{X}_k(\gamma)$, $\Sigma_k(\gamma)=S_k(\gamma)$, then we have

$$C'-\frac{1}{2}\sum_{k=1}^{g}|c_k(\gamma)|\log|S_k(\gamma)|$$

Here we have used

$$\begin{aligned}\sum_{i\in c_k}(X_i-\bar{X}_k)'S_k^{-1}(X_i-\bar{X}_k)&=\operatorname{tr}\left(\sum_{i\in c_k}S_k^{-1}(X_i-\bar{X}_k)(X_i-\bar{X}_k)'\right)\\&=\operatorname{tr}\Bigg(S_k^{-1}\Bigg[\underbrace{\sum_{i\in c_k}(X_i-\bar{X}_k)(X_i-\bar{X}_k)'}_{S_k}\Bigg]\Bigg)\end{aligned}$$

25.2. **k-means.** Choose γ to minimize

$$\sum_{k=1}^{g}\sum_{x_i\in c_k}\|X_i-\bar{X}_k\|^2$$

In Matlab, we write this as kmean($\underset{\sim}{X},g$). $\underset{\sim}{X}$ is $n\times p$.

25.3. **Hierarchical Clustering.**

- $Z=$ linkage($\underset{\sim}{X}$,'single')
  We can also put 'complete' or 'average'. $\underset{\sim}{X}$ is $n\times p$.
- $\gamma=$cluster(Z,'maxclust',g)
  γ is $n\times 1$.

25.3.1. *Procedure.*

- Suppose we have $x_1, \dots, x_n$
- For each pair x_i, x_j, we calculate the distance $d(x_i, x_j) =: d_{ij}$.
- Start with $c_i = \{x_i\}, 1 \le i \le n$, that is each point is in its own cluster.
- Merge: If
$$d_{ij} = \min_{s,t}\{d_{st}\}$$
then we merge $\{i, j\}$ to one cluster
-
$$d_{i',j'} = \min_{s,t}\{d_{st}\}$$
excluding $\{i, j\}$.
$$\{i', j'\} \cap \{i, j\} \begin{cases} = \emptyset & \text{merge}\{i', j'\} \\ \neq \emptyset & \text{merge}\{i, j\} \end{cases}$$
- We stop this when the inter-cluster distance exceeds a threshold $d_0 > 0$.

The difference: Suppose d_{12} is the smallest distance and the $\{1, 2\}$ are merged. For single linkage, when we calculate distance between $\{1, 2\}$ and j, we use $\min\{d_{ij}, d_{2j}\}$ and for complete linkage, we use $\max\{d_{1j}, d_{2j}\}$.

25.4. **Spectral Method.** Suppose we have Normal data
$$X_i \sim \mathcal{N}(\mu_k, \Sigma)$$
Then we have
$$\underbrace{\underset{\sim}{X}}_{n\times p} = \begin{bmatrix} X_1' \\ \vdots \\ X_n' \end{bmatrix} = M + \underset{\sim}{Z}$$
where
$$M = \underbrace{L}_{n\times g} \underbrace{\begin{pmatrix} \mu_1' \\ \vdots \\ \mu_g' \end{pmatrix}}_{g\times p}$$
$L_{ik} = 1$ if and only if $i \in c_k$. c_k true, that is each row of L has exactly one 1 and the rest are 0.

There are two options

- Use k means to $\underset{\sim}{X}$
- Use k means to SVD of $\underset{\sim}{X}$.
$$\underset{\sim}{X} = \hat{U}\hat{D}\hat{V}'$$
$$M = UDV'$$

U is $n \times p$ and D is $p \times p$, but $D = \text{diag}(d_1, d_2, \dots, d_g, 0, \dots, 0)$ if $g \le p$. We hope that the first g columns of $\hat{U}$ is approximately those of U. Then we apply k means to first g columns of $\hat{U}$.

We now calculate U. Columns of U are eigenvectors of MM' associated with the g nonzero eigenvalues of MM'. Let 1_k be the "unknown" vector
$$1_k(i) = \begin{cases} 1 & i \in c_k \\ 0 & \text{otherwise} \end{cases}$$
and
$$\theta_k = \frac{1_k}{\|1_k\|}$$
We have
$$L = [1_1, 1_2, \dots, 1_g] = [\theta_1 \cdots \theta_g] \begin{bmatrix} \|1_1\| & & \\ & \ddots & \\ & & \|1_g\| \end{bmatrix}$$

so

$$\begin{aligned} MM' &= L \begin{pmatrix} \mu_1' \\ \vdots \\ \mu_g' \end{pmatrix} (\mu_1 \cdots \mu_g) L' \\ &= [\theta_1 \cdots \theta_g] \begin{bmatrix} \|1_1\| & & \\ & \ddots & \\ & & \|1_g\| \end{bmatrix} \Big((\mu_i, \mu_j)_{1 \le i,j \le g} \Big) \begin{bmatrix} \|1_1\| & & \\ & \ddots & \\ & & \|1_g\| \end{bmatrix} \begin{bmatrix} \theta_1' \\ \vdots \\ \theta_g' \end{bmatrix} \\ &\equiv \underbrace{\Theta}_{n \times g} \underbrace{\tilde{A}}_{g \times g} \underbrace{\Theta'}_{g \times n} \end{aligned}$$

We wish to find eigenvectors of MM' associated with nonzero eigenvalues.

Our guess is

$$\eta = \sum_{k=1}^{g} a_k \theta_k = \Theta a$$

where

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_g \end{pmatrix}$$

which is $g \times 1$.

Now,

$$\begin{aligned} MM'\eta &= \underbrace{\Theta \tilde{A} \Theta'}_{MM'} \underbrace{\Theta a}_{\eta} \\ &= \Theta \tilde{A} a \end{aligned}$$

If η is an eigenvector associated with eigenvalue λ, then this equals

$$\begin{aligned} MM'\eta &= \lambda \eta \\ &= \lambda \Theta a \end{aligned}$$

η indeed is an eigenvector if $\tilde{A}a = \lambda a$.

Lemma. *The kth column of U, $1 \le k \le g$, has the form of*

$$\pm \sum_{i=1}^{g} a_i^{(k)} \theta_i = \pm \sum_{i=1}^{g} \frac{a_i^{(k)}}{\|1_i\|} 1_k$$

where $a^{(k)}$ is the kth leading eigenvector of $\tilde{A}$.

- Pull out the first g leading eigenvectors of $\underset{\sim}{X}\underset{\sim}{X}'$
- Form a $n \times g$ matrix
- Use k-means

Let us look at the case where $g = 2$

- $p \gg n$
- $g = 2$
- Also assume X_i are renormalized

$$\begin{aligned} X_i &= \mathcal{N}(-(1-\delta)\mu, \Sigma) \quad \text{class 1} \\ X_i &= \mathcal{N}(\delta\mu, \Sigma) \quad \text{class 2} \end{aligned}$$

Here, δ is the fraction of samples in class 1

The class labels now become ± 1. Let us say l, which is $n \times 1$ is such that

$$l_i = \begin{cases} -(1-\delta) & i \in \text{class 1} \\ \delta & i \in \text{class 2} \end{cases}$$

Then label

$$\gamma = \text{sgn}(l)$$

Now,

$$\underset{\sim}{X} = l \cdot \mu' + \underbrace{\underset{\sim}{Z}}_{\text{noise}}$$

When you use spectral clustering,

$$\underset{\sim}{X}\underset{\sim}{X}' = \underbrace{l(u'u)l'}_{\|u\|^2 ll'} + lu'Z + Zul' + ZZ'$$

The first leading eigenvector of $\underset{\sim}{X}\underset{\sim}{X}'$ is approximately equal to that of $\|u\|^2 ll' \propto l$

At least, you need

$$(\|u\|^2\|l\|^2) \gg \|ZZ'\| = O(\max(p,n))$$

The approach is to do feature selection first. We remove most features, leaving say $k \ll \min(p,n)$ only. Then use spectral clustering.