437 Midterm1.

Projection Matrix.
In the context of linear Algebra,
projection Matrix is the orthogonal projection onto the column space of X s.t. $P^2 = P$ (idempotent) and P is symmetric $P = P^T$ ↓

1. [8] Let $C = I_n - \frac{1}{n}1_n1_n^T$ be the centering operator on $\mathbb{R}^n$.

(a) [2] Show that C is positive semi-definite (PSD).

(a). 先证C是Projection Matrix ⇓ ProjectionMatrix是PSD的

Wts: $X^TCX \geq 0$, $\forall X \in \mathbb{R}^n$

We claim C satisfies that $C = C^T$ and $C^2 = C$.

For $C^T = (I - \frac{1}{n}1_n1_n^T) = I^T - \frac{1}{n}(1_n1_n^T)^T = I - \frac{1}{n}11^T = C$

And $C^2 = (I - \frac{1}{n}1_n1_n^T)(I - \frac{1}{n}1_n1_n^T) = I^2 - \frac{2}{n}1_n1_n^T + \frac{1}{n^2}1(1_n^T1)1_n^T = I^2 - \frac{2}{n}1_n1_n^T + n\cdot\frac{1}{n^2}1_n1_n^T$

$= I^2 - \frac{1}{n}1_n1_n^T = C$

So, C is a projection Matrix

For $X^TCX = X^TCCX = (X^TC^T)(CX) = (CX)^T(CX) = \|CX\|^2 \geq 0$

$\Rightarrow$ PSD.

(b) [2] Explain why the rank of C is $n-1$.

(b). $J = \frac{1}{n}1_n1_n^T$ 的投影Rank是为1 ⇓ $C = 1 - J$, J是一个projection MTX $\Rightarrow$ Rank(C) = n-1

① Orthogonal Complement:
For $S \subseteq \mathbb{R}^n$, $S^\perp$ is the orthogonal complement s.t. $S^\perp = \{s \in S : \langle s, s^\perp \rangle = 0, \forall s^\perp \in S^\perp\}$
$\dim(\mathbb{R}^n) = \dim(S) + \dim(S^\perp)$

② Projection Complement.
For projection Matrix P, let $Q = I - P$, Q is also a projection Matrix but projects to $S^\perp$ for P projects to S.

$C = I_n - \frac{1}{n}1_n1_n^T$ (I_n spans $\mathbb{R}^n$)
$= I_n - J$, let $J = uu^T = (\frac{1}{\sqrt{n}}1_n)(\frac{1}{\sqrt{n}}1_n^T) = \frac{1}{n}11^T$, For $\|u\| = 1$, in the form of $\frac{aa^T}{a^Ta} = P$ ↑ $= n$

(c) [2] Determine the eigenvalues of C and their multiplicities. (Recall: the multiplicity of an eigenvalue λ is the number of linearly independent eigenvectors corresponding to λ.)

(c) Projection MTX ⇓ idempotent ⇓ $\lambda = 1$ or 0 ⇓ Rank(n-1) $\Rightarrow \lambda = 1$ for n times

By (a), $C^2 = C \Rightarrow \lambda^2 = \lambda = 1$

As $Rank(C) = n-1$, and $\lambda = 1$ or 0,
we have 1's multiplicity is n-1.
0's multiplicity is 1.

We conclude $\frac{1}{n}11^T$ is projection MTRX that projects anything onto Span(1).
$\dim(\text{Image}(J)) = 1$
$\rightarrow \dim(\text{Image}(C)) = n - \dim(\text{Image}(J))$
$= n - 1 = n - 1$

(d) [2] Show that $C1_n = 0$, i.e., 1_n is an eigenvector corresponding to the zero eigenvalue.

手动计算后验证.

$C1_n = (I_n - \frac{1}{n}1_n1_n^T)1_n$

$= \begin{bmatrix} \cdots \\ \cdots \end{bmatrix}\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} - \frac{1}{n}1_n1_n^T1_n = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = 0 \Rightarrow C1_n = 0 \cdot 1$ ($1_n^T1_n = n$)

$\Rightarrow 1$ is the eigenvector of C with 0 to be its eigenvalue

$AV = \lambda V$ (V: 其特征向量 / 特征向量对应的特征值; A: 目标矩阵; λ: 特征值) 求特征向量的

SAMPLE MEAN

$\bar{x} = \frac{X^T 1_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$

SAMPLE COV

$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T = \frac{X^T C X}{n}$ for

C to be the centering operator. 对已center的, 则是 $\frac{X^T X}{n}$.

$\frac{X^T 1_n}{n} = \frac{1}{n}\begin{pmatrix} f_1^T \\ \vdots \\ f_p^T \end{pmatrix} 1_n = \begin{pmatrix} \bar{f}_1 \\ \vdots \\ \bar{f}_p \end{pmatrix}$

2. [8] Let $X = \begin{pmatrix} | & | & | \\ f_1 & f_2 & f_3 \\ | & | & | \end{pmatrix} = \begin{pmatrix} - & x_1^T & - \\ & \vdots & \\ - & x_{10}^T & - \end{pmatrix} \in \mathbb{R}^{10\times 3}$, be a data matrix where rows correspond to observations and columns correspond to variables. Assume that X is *column-centered* (each column has mean 0) and *column-orthogonal* ($X^T X = I_3$).

(a) [2] Find the sample mean vector $\bar{x}$ and the sample covariance matrix S of X.

Mean: $X = \begin{bmatrix} x_{11}-m_1 & x_{21}-m_2 & x_{31}-m_3 \\ x_{12}-m_1 & x_{22}-m_2 & \\ x_{13}-m_1 & x_{23}-m_2 & \\ \vdots & \vdots & \end{bmatrix} \Rightarrow \begin{bmatrix} \sum_{i=1}^{n} x_{1i}-m_1 & \sum_{i=1}^{n} x_{2i}-m_2 & \sum_{i=1}^{n} x_{3i}-m_3 \end{bmatrix}^T$

样本均值是每一列平均值相加之后的向量, 和行无关. 结果为 #column × 1 的 Vector.

$\bar{x} = \begin{bmatrix} \bar{f}_1 \\ \bar{f}_2 \\ \bar{f}_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \vec{0} \in \mathbb{R}^{3\times 1}$

VARIANCE.

$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{10} I_3 = \begin{bmatrix} \frac{1}{10} & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & \frac{1}{10} \end{bmatrix}$ 的 Vector.

(b) [2] Suppose f_1 represents measurements in meters, f_2 in centimeters, and f_3 in millimeters. Convert all measurements to centimeters by defining $Y = \begin{pmatrix} | & | & | \\ 100 f_1 & f_2 & 0.1 f_3 \\ | & | & | \end{pmatrix} \in \mathbb{R}^{10\times 3}$. What is the sample covariance matrix of Y?

$\begin{bmatrix} 100 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.1 \end{bmatrix} \Rightarrow 100.$

① Convertion. $X \; y \Rightarrow \left[\begin{array}{c|c|c} 100 f_{1,1} & f_{2,1} & 0.1 f_{3,1} \\ \vdots & \vdots & \vdots \\ 100 f_{1,10} & f_{2,10} & \end{array}\right] = Y$ (10×3 · 3×3)

$\begin{bmatrix} f_{1,1} & f_{2,1} & f_{3,1} \\ \vdots & \vdots & \vdots \end{bmatrix}\begin{bmatrix} 100 & & \\ & 1 & \\ & & 0.1 \end{bmatrix} = \left[\begin{array}{c|c|c} f_{1,1}\cdot 100 + f_{2,1}\cdot 0 + f_{3,1}\cdot 0 & f_{1,1}\cdot 0 + f_{2,1}\cdot 1 + f_{3,1}\cdot 0 & f_{1,1}\cdot 0 + f_{2,1}\cdot 0 + f_{3,1}\cdot 0.1 \\ f_{1,2}\cdot 100 + \cdots & \vdots & \vdots \end{array}\right]$

(c) [2] Let $v \in \mathbb{R}^3$ be a vector with unit norm, i.e., $\|v\| = 1$. Define $z = Xv$. Find the sample mean of z.

$z = Xv$, $v \in \mathbb{R}^3$, $\|v\| = 1$

Consider $Cz = CXv = \left[I_n - \frac{1}{n}1_n 1_n^T\right]Xv = z$

$= I_n z - \frac{1}{n}1_n 1_n^T z$ ($1_n^T z = 1_n^T Xv$)

$= z - 0 = z$.

$= (1^T X)v$

$= \vec{0}^T v = \vec{0}$ ([0 0 0] 1×3 MATRIX)

$\Rightarrow Cz = z$

$\Rightarrow z$ is column centered, z's mean is 0

② SAMPLE COV OF Y.

$\left[\frac{Y^T Y}{n}\right] = \left[\frac{DX^T X D}{n}\right] = \frac{1}{10}(D I_3 D) = \frac{1}{10}D^2$

$I_3 = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & \end{bmatrix}$

$= \begin{bmatrix} 1000 & & \\ & \frac{1}{10} & \\ & & \frac{1}{1000} \end{bmatrix}$

$\Rightarrow D^2$ Ans

(d) [2] What is the sample variance of z?

known centered

$S_z = \frac{z^T z}{n} = \frac{v^T X^T X v}{n}$

$= \frac{v^T v}{n} = \frac{\|v\|^2}{n}$ (note)

$= \frac{1}{10}$

Linear Transformation of MVN.

For $y \sim N_n(\mu, \Sigma)$, let $z = Ay + b$

then $z \sim N_m(A\mu + b, A\Sigma A^T)$

TRANSPOSE 矩阵操作: $(AB)^T = B^T A^T$, $(A^T)^T = A$

3. [16] Consider the linear regression model

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N_n(0, \sigma^2 I_n),$$

where $X \in \mathbb{R}^{n \times p}$ is a fixed *column-orthogonal* data matrix ($X^T X = I_p$), $\beta \in \mathbb{R}^p$ is a vector of regression coefficients, and ε is Gaussian noise. Assume that σ^2 is known.

(a) [2] What is the distribution of y? Write it down in terms of X, β, and σ^2.

$E(y) = E(X\beta + \varepsilon) = X\beta + E(\varepsilon) = X\beta$

↑ Constant in context $\Rightarrow E(\text{Constant}) = \text{Constant}$

$E(\varepsilon) = 0$ by Assumption

$\text{Var}(y) = \text{Var}(X\beta + \varepsilon) = 0 + \text{Var}(\varepsilon) = \sigma^2 I_n$ ← Constant Variance = 0

$\Rightarrow y \sim N_n(X\beta, \sigma^2 I_n)$

(b) [2] Note that for column-orthogonal X, the regression coefficient vector is $\hat{\beta} = X^T y$. Show that the fitted values $\hat{y} = X\hat{\beta}$ can be written as $\hat{y} = Py$, where P is a projection matrix. Specify the space onto which P projects.

Fitted Value: $\hat{y} = X\hat{\beta}$

$\hat{y} = X\hat{\beta} = X(X^T y) = (XX^T)y = Py$

For $\hat{\beta}$, Use Least Squares Estimates.

$\hat{\beta} = (X^T X)^{-1}(X^T y) = (I_p)^{-1}(X^T y) = X^T y$

↑ Column Orthogonal

For $(XX^T)(XX^T) = X(X^T X)X^T = P$ ($X^T X = I_p$) $\Rightarrow$ Idempotent

And $P^T = (XX^T)^T = (X^T)^T X^T = XX^T = P$ $\Rightarrow$ Symmetric

$(X^T)^T X^T = XX^T$

both satisfies $\Rightarrow$ P

MVN & Linear Transformation

(c) [2] What is the distribution of $\hat{y}$?

$y \sim N_n(X\beta, \sigma^2 I_n)$

For $E(\hat{y}) = PE(y)$ from (b). For (column ortho-X) $X^T X = I_p$

$E(\hat{y}) = P(X\beta) = (XX^T)(X\beta) = X\beta = E(y)$

$A\mu + b|_{b=0, A=P}$

(d) [2] Show that the residuals $r = y - \hat{y}$ can be expressed as $r = P_\perp y$ for some projection matrix $P_\perp$. Specify the space onto which $P_\perp$ projects.

For $\text{Var}(\hat{y}) = P(\text{Var}(y))P^T$ ↓ MVN variance

$A\Sigma A^T$, $A = P$ here.

$= P(\sigma^2 I_n)P^T = \sigma^2(PI_nP^T) = \sigma^2 PP^T = \sigma^2 P^2 = \sigma^2 XX^T$

($P^2 = P$, $P^T = P$)

So $\hat{y} \sim N_n(X\beta, \sigma^2 XX^T)$

$C(X)$: column space: X所有列向量可组成的空间

(d): RESIDUAL = OBSERVED − FITTED = $y - \hat{y} = r$

(Response, Response)

I_n 是 introduced, 用于提取出来, ($I_n y = y$)

$r = y - \hat{y} = I_n y - Py = (I_n - P)y$, For $P = XX^T$, $r = (I_n - XX^T)y \Rightarrow P_\perp = (I_n - XX^T)$

$P_\perp$ projects onto $C(X_\perp)$

q1 有说, 正交补空间的投影

① $XX^T = P \Rightarrow I - XX^T = P_\perp$ ②

(互为正交补 ORTHOGONAL COMPLEMENT)

(e) [2] What is the distribution of the residuals r?

$E(r) = E(P_\perp y) = E[(I_n - P)y] = (I_n - XX^T)E(y) = (I_n - XX^T)(X\beta) = X\beta - XX^TX\beta$ (I)

$= X\beta - X\beta = \vec{0}$

$Var(r) = (I_n - P)\, Var(y)\, (I_n - P)^T = (I_n - P)(\sigma^2 I_n)(I_n - P)^T$ ($= (I_n - P)$)

$= \sigma^2 (I_n - P)^2$ → idempotent

$= \sigma^2 (I_n - P) = \sigma^2 (I_n - XX^T)$

(MATRIX FORM: $Var(Ay) = A\, Var(y)\, A^T$)

So $r \sim N_n(0, \sigma^2(I_n - XX^T))$

(f) [2] Explain why the residuals r and the fitted values $\hat{y}$ are independent. (MVN.)

r and $\hat{y}$ independent

⇅

WTS: $Cov(r, \hat{y}) = 0$

$Cov(Ay, By) = A\, Var(y)\, B^T$

(UNIVARIATE) $Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$

$Cov(\hat{y}, r) = Cov(Py, P_\perp y) = P\, Var(y)\, P_\perp = P(\sigma^2 I) P_\perp$

$= \sigma^2 P(I - P) = 0$ ($PP_\perp = 0$) ⇒ $\hat{y}$ and r are uncorrelated, thus independent.

(g) [2] Write down the joint distribution of $\begin{pmatrix} \hat{y} \\ r \end{pmatrix}$.

$$\begin{pmatrix} \hat{y} \\ r \end{pmatrix} \sim N_{2n}\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} XX^T & 0 \\ 0 & I_n - XX^T \end{pmatrix}\right)$$

(h) [2] What is the conditional distribution of $r|\hat{y}$?

r and $\hat{y}$ are independent

$\Rightarrow r|\hat{y} \sim N_n(0, \sigma^2(I - XX^T))$

[END OF EXAMINATION; total points = 32]

MIDTERM 2

① Multivariate Normal Distribution
② MVN Linear Transformation
③ Matrix Calculus.

① $f(y) = (2\pi)^{\frac{-n}{2}} |\Sigma|^{-\frac{1}{2}} e^{(-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu))}$

② For constant $A \in \mathbb{R}^{p\times n}$ and $y \sim N_n(\mu, \Sigma)$, $z = Ay \sim (A\mu, A\Sigma A^T)$

③ For vector β, $\beta^T A \beta$'s derivative, $\partial_\beta[\beta^T A\beta]$ is $2A\beta$.

1. [8] Consider the linear regression model

$$y = X\beta + \varepsilon,$$

where $X = \begin{pmatrix} -x_1^T- \\ \vdots \\ -x_n^T- \end{pmatrix} \in \mathbb{R}^{n\times p}$ is a fixed *column-orthogonal* design matrix ($X^TX = I_p$), $\beta \in \mathbb{R}^p$ is a vector of unknown regression coefficients, and $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$ is a Gaussian noise vector. Assume that σ^2 is fixed and known. From Midterm 1, we know that the response vector follows a multivariate normal distribution:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \sim \mathcal{N}_n(X\beta, \sigma^2 I_n).$$

(a) [2] Write down the **log-likelihood** function $\ell(y_1, \ldots, y_n; \beta)$, and explain why finding the maximum likelihood estimate (MLE) for β is equivalent to solving the least-squares problem, i.e., minimizing $\|y - X\beta\|^2$ with respect to β.

MLE ⇆ Least Squares. Goal of optimization is the same

For model assumption, $y \sim N_n(X\beta, \sigma^2 I_n)$, $\mu = X\beta$ and $\Sigma = \sigma^2 I_n$, $\det\Sigma = |\sigma^2 I_n| = (\sigma^2)^n$

$f(y;\beta) = (2\pi)^{\frac{-n}{2}} (\sigma^2)^{\frac{-n}{2}} e^{(-\frac{1}{2}(y-X\beta)^T (\frac{1}{\sigma^2} I_n)(y-X\beta))}$ with $\Sigma^{-1} = (\sigma^2 I_n)^{-1} = \frac{1}{\sigma^2} I_n$

$\ell(y;\beta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\|y - X\beta\|^2$

$\ln(2\pi)$: constant; $\ln(\sigma^2)$: Assumption constant; $\|y-X\beta\|^2 \geq 0$

⇒ Maximize $(-\|y - X\beta\|^2)$ to maximize ℓ

⇓

Minimize $\|y - X\beta\|^2$ in OLS. □

TRANSPOSE
$(A-B)^T = A^T - B^T$
$(AB)^T = B^TA^T$

(b) [2] From the previous part, we conclude that the MLE is $\hat\beta = X^Ty$. What is the distribution of $\hat\beta$?

$\partial_\beta[(y-X\beta)^T(y-X\beta)] = \partial_\beta[y^Ty - 2y^TX\beta + \beta^T(X^TX)\beta]$ ⇒ $2I_p\beta = 2y^TX$

$(y-X\beta)^T = y^T - (X\beta)^T = (y^T - \beta^TX^T)$

$-2y^TX\beta$; $\beta^T(X^TX)\beta \Rightarrow 2(X^TX\beta)$; $X^TX = I_p$

$\Rightarrow \beta = y^TX$ (Final Result)

$= y^Ty - y^TX\beta - \beta^TX^Ty + \beta^TX^TX\beta$ (SAME!)

Trans pose: $(y^TX\beta)^T = \beta^TX^Ty$

VARIANCE MTRX FORM

$Var(Ay) = A\,Var(y)\,A^T$

y: $n\times 1$, X: $n\times p$, β: $p\times 1$ ⇒ 1×1 For $y^TX\beta$ ⇒ Result is scaler

Transpose of scaler is still scaler ⇒ Merge into 1.

$\hat\beta = X^Ty$

$E(\hat\beta) = E(X^Ty) = X^TE(y) = X^TX\beta = I_p\beta = \beta$

⇒ $\hat\beta$ is unbiased estimator.

$Var(\hat\beta) = Var(X^Ty) = X^T(Var(y))X = X^T\sigma^2 I_n X = \sigma^2 X^TX = \sigma^2 I_p$

⇒ $\hat\beta \sim N_p(\beta, \sigma^2 I_p)$ Result.

HYPOTHESIS TESTING // 1. test statistic in context. MONO $\mathbb{R}$ | MULTI $\mathbb{R}^n$

$\mathbb{R}$: $z_j = \frac{\hat\beta_j - 0}{\sqrt{\sigma^2}} = \frac{\hat\beta_j}{\sigma} \sim N(0,1)$

$z_j^2 = \left(\frac{\hat\beta_j}{\sqrt{\sigma^2}}\right)^2 = \hat\beta_j (\sigma^2)^{-1} \hat\beta_j$

$\mathbb{R}^n$: 依附 $D^2 = (x-\mu)^T \Sigma^{-1} (x-\mu)$

test statistic in context: $(\hat\beta - 0)^T (\sigma^2 I_p)^{-1} (\hat\beta - 0)$

$\Rightarrow \frac{1}{\sigma^2} \hat\beta^T I_p \hat\beta = \frac{\|\hat\beta\|^2}{\sigma^2} = \sum_{i=1}^{p} \left(\frac{\hat\beta_i}{\sigma}\right)^2$

(c) [2] Suppose you would like to test the hypothesis

$$H_0 : \beta = 0 \quad \text{vs.} \quad H_1 : \beta \neq 0.$$

Compute the test statistic corresponding to this hypothesis and indicate what distribution this statistic follows under the null hypothesis.

(这部分在 Multivariate Testing 里, MVN note)

$$f^2 = (\beta - 0)^T (\sigma^2 I_p)^{-1} (\hat\beta - 0) = \frac{\|\hat\beta\|^2}{\sigma^2} \sim f^2(p)$$

(Chi-square distribution) ← d.f. = p

test statistic = $\frac{\text{observed Bias}}{\text{Environmental Noise}}$

(d) [2] Suppose the p-value you obtained is less than 0.05. What can we conclude about the components of $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$?

p-value < 0.05 $\Rightarrow$  $\Rightarrow$ landed on Rejection Area

$\Downarrow$

Reject H_0 $\Leftrightarrow$ At least one $\beta_i \neq 0$ — Reject H_0

Eldog3 - Mixed Signals kurtosis will change: $y_1 \sim (\mu_1, \sigma_1^2)$, $y_2 \sim (\mu_2, \sigma_2^2)$, w.i.s. $\kappa(y_1+y_2) = \frac{\sigma_1^4 \kappa(y_1) + \sigma_2^4 \kappa(y_2)}{(\sigma_1^2+\sigma_2^2)^2}$

$\kappa(a_1 z_1 + a_2 z_2) = a_1^4 \kappa(z_1) + a_2^4 \kappa(z_2)$. Assume y_1 and y_2 are independent.

① Centering: kurtosis is translation invariant, i.e. $\kappa(y) = \kappa(y-\mu)$, we can assume $\mu_1 = 0$, $\mu_2 = 0$ WLOG.

Let $\tilde y_i = y_i - \mu_i$, we have $\tilde y_i$'s mean = 0 and variance is $\sigma_{sum}^2 = \sigma_1^2 + \sigma_2^2$ by independence. For $S = y_1 + y_2$,

By definition, $\kappa(y) = E\left[\frac{(y-\mu)^4}{\sigma}\right] - 3$ and for zero-mean variables, $\tilde\kappa = \kappa(\tilde y) = \frac{E(\tilde y^4)}{(E(\tilde y^2))^2} - 3 = \frac{E(S^4)}{E^2(S)}$

分子与分母本质是在对 Varian 进行操作.

By Expanding $(y_1+y_2)^4$, it gives $y_1^4 + 4y_1^3 y_2 + 6y_1^2 y_2^2 + 4y_1 y_2^3 + y_2^4$. $\Rightarrow E(\text{Polynomial}) = E(y_1^4) + 4E(y_1^3 y_2) + E(6y_1^2 y_2^2) + E[4y_1 y_2^3] + E(y_2^4) = E(y_1^4) + 6\sigma_1^2\sigma_2^2 + E(y_2^4)$

$E[4y_1 y_2^3] = 0$

For y, we have $E(y^4) = (\kappa(y) + 3)\sigma^4$

So $E[(y_1+y_2)^2] = (\kappa_1\sigma_1^4 + 3\sigma_1^4) + 6\sigma_1^2\sigma_2^2 + (\kappa_2\sigma_2^4 + 3\sigma_2^4)$

$= \kappa_1\sigma_1^4 + \kappa_2\sigma_2^4 + 3(\sigma_1^4 + \sigma_2^4)^2$ ← 使用完全平方公式化简

$$\kappa(S) = \frac{\kappa_1\sigma_1^4 + \kappa_2\sigma_2^4 + 3(\sigma_1^2+\sigma_2^2)}{(\sigma_1^2+\sigma_2^2)^2} - 3 = \frac{\kappa_1\sigma_1^4 + \kappa_2\sigma_2^4}{(\sigma_1^2+\sigma_2^2)^2} + 3 - 3 = \frac{\sigma_1^4\kappa(y_1) + \sigma_2^4\kappa(y_2)}{(\sigma_1^2+\sigma_2^2)^2}.$$ □

MTRX

row : data point number

column : dimension of space

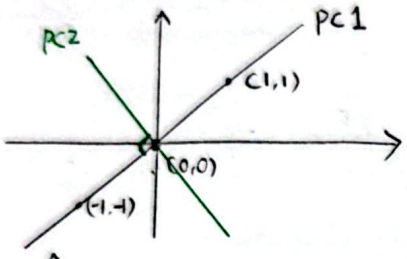
Row : data point → (0,0), (1,1), (-1,-1) } ⇒ Flattened into 2D plane

2. [8] Consider the following dataset with 3 observations and 2 features:

$$X = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ -1 & -1 \end{pmatrix} \in \mathbb{R}^{3\times 2}.$$

原始空间为 $\mathbb{R}^2$，数据点为 3

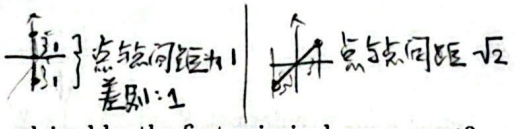
(a) [2] Find the first principal component direction v_1. (Hint: drawing a scatterplot may help.)



$\vec{v}_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$ for unit norm $\vec{v}_1$ (斜度)

只有在线上，三点的分布不会有改变，若不在 PC1，则：

(从 projection 的角度去思考)



(b) [2] How much variance is explained by the first principal component?

Variance Explained by PC: projection (dot product) 将原始数据投影至新坐标轴之上.

$z_{i(1)} = x_i^T v_1$: Variance explained by PC1

$$\begin{cases} 0\cdot\frac{1}{\sqrt{2}} + 0\cdot\frac{1}{\sqrt{2}} = 0 \\ 1\cdot\frac{1}{\sqrt{2}} + 1\cdot\frac{1}{\sqrt{2}} = \sqrt{2} \\ -1\cdot\frac{1}{\sqrt{2}} + (-1)\cdot\frac{1}{\sqrt{2}} = -\sqrt{2} \end{cases} \Rightarrow \bar{z}_{(1)} = 0+\sqrt{2}+(-\sqrt{2}) = 0 \Rightarrow \lambda_1 = \mathrm{Var}(z_{(1)}) = \frac{1}{3}\sum (z_i - \bar{z})^2$$

$$z_{(1)} = \begin{bmatrix} 0 \\ \sqrt{2} \\ -\sqrt{2} \end{bmatrix} \in \mathbb{R}^3 \qquad = \frac{1}{3}\sum (z_i)^2 = \frac{1}{3}(0+2+2) = \frac{4}{3}$$

(c) [2] Find the second principal component direction v_2. How much variance is explained by the second component?

$v_2 \perp v_1 \Rightarrow \langle v_2, v_1 \rangle = 0 \Rightarrow v_2 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ ← (对称性取负)

$\bar{z}_{(2)} = 0+0+0 = 0 \Rightarrow \lambda_2 = \mathrm{Var}(z_2) = \frac{1}{3}\sum (z_i - 0)$

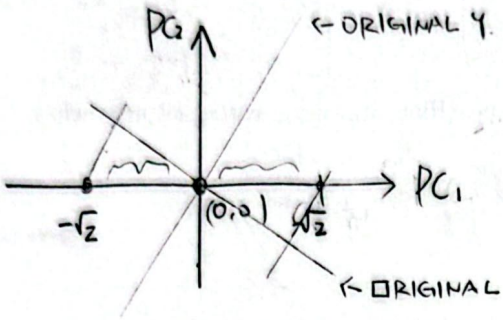
$z_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ $= 0 \Rightarrow$ Variance Explained is 0.

(b) 更直接的说明

$K(a,b) = \langle a, b\rangle \Rightarrow \varphi(a) = a,\ \varphi(b) = b \Rightarrow$ No Transformation Applied.

$= \langle \varphi(a), \varphi(b)\rangle$ (in context)

(d) [2] Draw a biplot using the first and second principal components, and indicate the **proportion of variance explained** by each component.



← ORIGINAL Y.

3 points has no projection on PC_2.

$PVE_1 = \frac{PC_1 \text{ Variance}}{\text{Total Variance}} = \frac{\frac{4}{3}}{\frac{4}{3}+0} = 1$

← ORIGINAL X-AXIS

$PVE_2 = \frac{0}{\frac{4}{3}+0} = 0$

3. [4] Suppose we apply kernel PCA to a dataset $X = \begin{pmatrix} -x_1^T- \\ \vdots \\ -x_n^T- \end{pmatrix} \in \mathbb{R}^{n\times p}$ using a kernel function $\mathcal{K}(a,b)$. We compute the kernel matrix $K \in \mathbb{R}^{n\times n}$ with entries $K_{ij} = \mathcal{K}(x_i, x_j)$.

(a) [2] Is the kernel matrix K always positive semidefinite? Explain why or why not.

Kernal PCA: φ: New Feature Space Coordinate

K: Gram Matrix.

Matrix A is PSD if for $\|\vec{v}\| \neq 0, \forall \vec{v} \in \mathbb{R}^n$ we have $\vec{v}^T A \vec{v} \geq 0$

For φ, let $\varphi: \mathbb{R}^p \to \mathbb{R}^d$, then $K_{ij} = \varphi(x_i)^T \varphi(x_j)$ $K = \varphi\varphi^T$ So $\vec{v}^T K \vec{v} = \vec{v}^T \varphi \varphi^T \vec{v} = \|\varphi^T \vec{v}\|^2 \geq 0$

$K = \varphi\varphi^T$ 所有 dot product 组成的 GRAM MATRIX

$\forall v \Rightarrow$ PSD

(b) [2] Explain why using the linear kernel $\mathcal{K}(a,b) = \langle a, b\rangle$ in kernel PCA is equivalent to performing standard PCA on the original dataset.

(SVD) ⇓

$A = UDV^T$ (分解) $\uparrow (V^T, V^TV = I)$

Linear Kernal PCA = Standard PCA ; Proof.

For $K(a,b) = \langle a,b\rangle = a^Tb \Rightarrow \varphi(x) = x$

$K_{ij} = x_i^T x_j \Rightarrow K = XX^T \in \mathbb{R}^{n\times n}$, Goal: $(XX^T)\lambda = \lambda\lambda$, For XX^T, SVD by doing

SVD 适用于任何矩阵.

$S \propto X^TX$, Goal: $(X^TX)w = \lambda w$

D: 对角线上是 σ_i, 其它为 0 的 diagonal matrix

U: 正交矩阵 $\in \mathbb{R}^{n\times n}$, $U^TU = I_n$

Sample Covariance Matrix

For $X = UDV^T$ by SVD, $X^TX = (VDU^T)(UDV^T) = VD^2V^T$

$(UDV^T)(VDU^T) = UD^2U^T$

⇒ Goal is Same As $X^TX = XX^T$ in t

(SVD 与 SD 相同, 后者仅适用于方阵 $C_{n\times n}$ ↳ Spectral (eigen) decomp)

$(\varphi^T v)^T$

$$\begin{pmatrix} | & | \\ f_1 & f_2 \\ | & | \end{pmatrix}\begin{pmatrix} - & f_1^T & - \\ - & f_2^T & - \end{pmatrix} \Rightarrow \left[f_1 f_1^T + f_2 f_2^T \right]$$

4. [6] Consider two centered feature columns $f_1 \in \mathbb{R}^n$ and $f_2 \in \mathbb{R}^n$ that are uncorrelated and have unit sample variance:

指的就是 $\rightarrow$ $s_{f_1}^2 = s_{f_2}^2 = 1, \quad s_{f_1 f_2} = 0.$

$\Downarrow$ $Cov(f_1, f_2) = 0$

We observe two datasets: $Var(f_1) = 1$ 另一个同理

$$X = \begin{pmatrix} | & | \\ f_1 & f_2 \\ | & | \end{pmatrix} \in \mathbb{R}^{n \times 2}, \qquad Y = \begin{pmatrix} | \\ f_1 \\ | \end{pmatrix} \in \mathbb{R}^{n \times 1}.$$

2个feature

(a) [2] Write down the sample covariance matrices S_X, S_Y, and S_{XY} required for canonical correlation analysis.

NOTATION: S_x: x's sample Covariance Matrix

S_{xy}: x and y's cross covariance Matrix

$$S_X = \begin{pmatrix} Var(f_1) & Cov(f_1, f_2) \\ Cov(f_2, f_1) & Var(f_2) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{2 \times 2}$$

$$S_y = 1 \in \mathbb{R}$$

$$S_{xy} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \in \mathbb{R}^{2 \times 1}$$

有关Block表达与Scaler/Input表达相互间的转译以及更多，看M2 Q4的解析

(b) [2] What is the first canonical correlation between X and Y? (Hint: you can solve this question without computing an SVD, just use the structure of the data.)

(有关First/Second的解释与区分，见M2Q4解析)

Ans: CCA finds linear combination of columns X (i.e. f_1 and f_2) that are max correlated with y (i.e. f_1)

Since y coincides with the 1st column of X

We can make this correlation = 1

i.e. $U = 1 \cdot f_1 + 0 \cdot f_2$

$V = 1 \cdot f_1$

Max Corr between X, Y

$\Downarrow$

如何让X与Y最像？ $\Rightarrow$ 因为X里有Y 那丢弃Y的话就可以

(c) [2] Determine the corresponding canonical directions u_1 for X and v_1 for Y.

For X, we only ~~seek for~~ keep first column and omit second.

So $U_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ meaning: $\in \mathbb{R}^{2 \times 1}$

For Y, we keep everything $\rightarrow$ $V_1 = 1 \in \mathbb{R}$

To verify Subject to constraints of $U_1^T \Sigma_x U_1 = V_1^T \Sigma_y V_1 = 1$

$Var(Xu_1) = 1$ As $U_1^T S_x U_1 = (1\ 0)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1$

$Var(XV_1) = 1$ As $V_1^T S_y V_1 = 1 \cdot 1 \cdot 1 = 1$

$\Rightarrow U_1^T S_{xy} V_1 = (1\ 0)\begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1$

① FACTOR ANALYSIS

MODEL DEFINITION

$X = \mu + LZ + \varepsilon$, $\mu = 0$

$\Rightarrow X = LZ + \varepsilon$.

$Z \sim N_r(0, I_r)$ — $Z \in \mathbb{R}^{r\times 1}$

$\varepsilon \sim N_p(0, \Psi)$ — $\varepsilon \in \mathbb{R}^{p\times 1}$, $L \in \mathbb{R}^{p\times r}$

ε and Z are independent

$X \in \mathbb{R}^{p\times 1}$

CONDITIONAL & JOINT NORMAL DISTRIBUTION ②

For $\begin{pmatrix} X \\ Z \end{pmatrix} \sim N_{p+r}\left(\begin{pmatrix} \mu_X \\ \mu_Z \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ} \end{pmatrix}\right)$,

↑ Joint. We have $X|Z \sim N(E(X|Z), Var(X|Z))$

With ① $E(X|Z) = \mu_X + \Sigma_{XZ}\Sigma_{ZZ}^{-1}(Z - \mu_Z)$ (固定公式) ② $Var(X|Z) = \Sigma_{XX} - \Sigma_{XZ}\Sigma^{-1}\Sigma_{ZX}$

Q1 LOGICFLOW: Marginal → Joint → Conditional

↳ $E(X) = E(LZ+\varepsilon) = LE(Z) + E(\varepsilon) = 0 + 0 = 0$

$Var(X) = Var(LZ+\varepsilon) = L\,Var(Z)L^T + Var(\varepsilon) = LL^T + \Psi$

↳ I_r

Note 9

1. Explain why factor model is equivalent to

$z \sim N(0, I)$

$\Downarrow$

$x|z \sim N(Lz, \Psi)$

Joint: For $Cov(X, Z)$

$= E[(X-\mu_X)(Z-\mu_Z)^T] = E(XZ^T) = E((LZ+\varepsilon)Z^T) = E(LZZ^T + \varepsilon Z^T)$ ($\mu_X = 0$, $\mu_Z = 0$)

$= LE(ZZ^T) + E(\varepsilon Z^T)$

$= L(I) + E(\varepsilon)E(Z^T)$

$= L \cdot I = L$ (↳ by independence Assumption)

$E(ZZ^T) = \begin{bmatrix} E(Z_1^2) & E(Z_1Z_2) \\ E(Z_2Z_1) & E(Z_2^2) \end{bmatrix}$

For $Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$, as an example.

Here, in context, $Z \sim N(0, I)$

So $E(Z_1) = 0 = E(Z_2)$

$Var(Z_1) + E^2(Z_1) = E(Z_1^2) = I$ ($E^2(Z_1) = 0$)

So $\begin{pmatrix} X \\ Z \end{pmatrix} \sim N_{p+r}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{pmatrix} LL^T+\Psi & L \\ L & I_r \end{pmatrix}\right)$

Consider distribution of $X|Z$.

$E(X|Z) = \mu_X + \Sigma_{XZ}\Sigma_{ZZ}^{-1}(Z - \mu_Z)$

$= 0 + L\,I^{-1}(Z - 0) = LZ$.

$Var(X|Z) = \Sigma_{XX} - \Sigma_{XZ}\Sigma^{-1}\Sigma_{ZX} = (LL^T + \Psi) - L\,I^{-1}L^T$

$= LL^T + \Psi - LL^T = \Psi$

$X|Z \sim N(LZ, \Psi)$

So we conclude $Z \sim N(0, I) \Rightarrow X|Z \sim N(LZ, \Psi)$

2. Show that PVE by j-th factor in FA via PCA is $\frac{\lambda_j}{tr(\Sigma)}$ where λ_j is j-th eval of Σ.

3. Is $PVE = \frac{\|L\|_F^2}{tr(\Sigma)}$ rotation invariant? I.e. will you get the same value of PVE if use $\tilde{L} = LQ$ instead? Are $PVE_j = \frac{\sum_{i=1}^{r}\ell_{ij}^2}{tr(\Sigma)}$ rotation invariant?

4. FA Solution is not guaranteed to exist. Consider:

$$\Sigma = \begin{pmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{pmatrix} \$for\$ x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

We want to use $r = 1$, i.e. find $L = \begin{pmatrix} \ell_{11} \\ \ell_{21} \\ \ell_{31} \end{pmatrix} \in \mathbb{R}^{3\times 1}$ and $\Psi = \mathrm{diag}(\psi_1, \psi_2, \psi_3)$ Such that $\Sigma = LL^T + \Psi$.

- Show that $\ell_{11}\ell_{21} = 0.9, \ell_{21}\ell_{31} = 0.4, \ell_{11}\ell_{31} = 0.7$
- Find value of ℓ_{11}
- Show that $\ell_{11} = \mathrm{Cor}(x_1, z_1)$ and explain why there is no solution to $\Sigma = LL^T + \Psi$
- Find value of ψ_1
- Show that it's not possible for $\psi_1 = \mathrm{var}(\epsilon_1)$

5. Finding scores via conditional distribution. Consider $y = \begin{pmatrix} x \\ z \end{pmatrix}$
   - Find joint distribution for y
   - Find conditional distribution $z|x$
   - Argue how to use formula for $E(z|x)$ to find scores $z_1 \dots z_n$
   - Why $E(z|x)$ is a good estimate?

Note 10

1. Assume that $P \in \mathbb{R}^{p\times p}$ is a permutation matrix that permutes i-th and j-th elements in $\mathbb{R}^p$, i.e.

Q2. ③ FA via PCA.

By spectral decomposition, covariance MTRX could be rewritten as: $\Sigma = \sum_{i=1}^{p}\lambda_i v_i v_i^T$ For (λ_i, v_i) to be the eigenvalue and eigenvector.

By selecting r number of leading eigenvectors & its corresponding eigenvalues, we have $\ell_i = \sqrt{\lambda_i}\, v_i$

Q3. ④ VARIMAX CRITERION.

↳ introducing orthogonal MTRX Q s.t. $QQ^T = I$

So $\tilde{L} = LQ$

Purpose: Make as many value close to 0 as possible. Resulting few large loadings.

RowVector: 行向量 $\in \mathbb{R}^{1\times n}$ ColumnVector: 列向量 $\in \mathbb{R}^{n\times 1}$

Q1. Assume P is a permutation matrix that permute i^{th} and j^{th} elements in $\mathbb{R}^p$.

i.e.: If $x = \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_j \\ \vdots \\ x_p \end{pmatrix}$ then $Px = \begin{pmatrix} x_1 \\ \vdots \\ x_j \\ \vdots \\ x_i \\ \vdots \\ x_p \end{pmatrix}$

PermutationMatrix: $n\times n$ square matrix.
Used switching row when PX, switching column when $X^T P$
All permutation matrix are orthonormal Matrix, as
$PP = I \Rightarrow P = P^T = P^{-1}$, meaning P is involutory matrix.

$P_{kk} = 1$ for $k \neq i$ or j
$\begin{cases} P_{ij} = P_{ji} = 1 \\ P_{ii} = P_{jj} = 0 \end{cases}$

a. Write down P explicitly

b. Write down P^{-1} explicitly

(Hint: note that $P^{-1}(Px) = P^{-1}\begin{pmatrix} x_1 \\ \vdots \\ x_j \\ \vdots \\ x_p \end{pmatrix} = x = \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_p \end{pmatrix}$)

ICA MODEL: $X = L \cdot Z$
$X \in \mathbb{R}^p$: MixureVector; $L \in \mathbb{R}^{p\times p}$: Unknown Mixing matrix; $Z \in \mathbb{R}^p$: Unknown SourceVector

ICA Goal: Find W s.t.
$W = L^{-1}$
$Z = WX$
to recover original Matrix

c. If $z = \begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix}$ and $L = (\ell_1 \dots \ell_p)$ write $\tilde{z} = Pz$ and $\tilde{L} = LP^{-1}$ in terms of L and z

d. Using equation $Lz = \ell_1 z_1 + \cdots + \ell_p z_p$ show that $\tilde{L}\tilde{z} = Lz$

2. Uncorrelated $\not\Rightarrow$ independent. Consider random vector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ with joint distribution

$$x = \begin{cases} (0,1) \\ (0,-1) \\ (1,0) \\ (-1,0) \end{cases} \text{ with probability } 1/4$$

- Show that $\operatorname{cor}(x_1, x_2) = 0$
- Find marginal distributions $f_{x_1}(x_1), f_{x_2}(x_2)$
- Show that x_1 and x_2 are not independent i.e. $f_x(x_1, x_2) = f_{x_1}(x_1) \cdot f_{x_2}(x_2)$

3. If $y_1 \sim (\mu_1, \sigma_1^2)$ and $y_2 \sim (\mu_2, \sigma_2^2)$ derive that

$$\mathcal{K}(y_1 + y_2) = \frac{\sigma_1^4 \mathcal{K}(y_1) + \sigma_2^4 \mathcal{K}(y_2)}{(\sigma_1^2 + \sigma_2^2)^2}$$

$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ 向量 ↑ scaler

Question 1. (a). $P = \begin{pmatrix} 1 & 0 & \cdots \\ & 0 & 1 \\ & 1 & 0 \\ & & & 1 \end{pmatrix}$ ← ith, jth.

define $P \in \mathbb{R}^{n\times n}$ for $P_{kk} = 1$ for $k \neq i$ or j
and $P_{ij} = P_{ji} = 1$, with $P_{ii} = P_{jj} = 0$

(b). Consider P^{-1}, since $P(PX) = X$
$\Rightarrow P^2 = I$
$\Rightarrow P = P^{-1}$ □

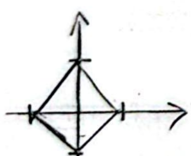
(c). For $\tilde{Z} = PZ$, $L \in [\,]^T$ (column vector)
$= P \cdot \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}$
$= \begin{bmatrix} z_1 \\ \vdots \\ z_j \\ \vdots \\ z_i \\ \vdots \\ z_n \end{bmatrix}$

For $\tilde{L} = LP^{-1} = (\ell_1, \ell_i, \ell_j, \ell) P^{-1} = (\ell_1 \dots \ell_j \dots \ell_i \dots \ell$
i^{th} j^{th} (SWAPPED) j^{th} i^{th}

(d). For $\tilde{L}\tilde{Z}$,
We have $[\ell_1 \cdots \ell_j \cdots \ell_i \cdots \ell_p] \begin{bmatrix} z_1 \\ \vdots \\ z_j \\ \vdots \\ z_i \\ \vdots \\ z_p \end{bmatrix} = \sum_{k=1}^{p} \ell_k z_k = LZ$.

Since $\sum_{k\neq i,j} (\ell_k + z_k) + (\tilde{\ell}_i \tilde{z}_i) + (\tilde{\ell}_j \tilde{z}_j)$
$= \sum_{k\neq i,j} (\ell_k + z_k) + (\ell_i z_i) + (\ell_j z_j)$
$= \sum_{k=1}^{p} \ell_k z_k = LZ$.

Q2. Uncorrelated $\not\Rightarrow$ independent. For $E(X) = \sum_{\text{All Possible } X} X\text{'s value} \cdot \text{Probability}$ ($\frac{1}{4}$ in Cont.)



By the question setup, following the definition of X,
We have: $E(X_1) = 0 \cdot \frac{1}{4} + 0 \cdot \frac{1}{4} + (-1) \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} = 0$
$E(X_2) = 1 \cdot \frac{1}{4} + (-1) \cdot \frac{1}{4} + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 0 = 0$
And $E(X_1 X_2) = 0$ For $X_1, X_2 \in \{0\cdot 1, (-1)\cdot 0, 1\cdot 0, (-1)\cdot 0\} = \{0, 0, 0, 0\}$

So $Cov(X_1, X_2) = E(X_1 X_2) - E(X_1)E(X_2) = 0 - 0 = 0$ $\Rightarrow X_1$ and X_2 is uncorrelated.

By definition of independence, if X_1 and X_2 are independent, then $f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$.

So PCA based on correlation is inefficient for signal seperation.

Here, $P(X_1 = 0) = P(0,1) + P(0,-1) = \frac{1}{2}$
Similarly, $P(X_2 = 0) = \frac{1}{2}$ $P(X_1 = 1) = \frac{1}{4}$ $P(X_2 = 1) = \frac{1}{4}$ $P(X_1 = -1) = \frac{1}{4}$, $P(X_2 = -1) = \frac{1}{4}$.
For $X = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $P(X = (0,1)) = \frac{1}{4}$, $f_{X_1}(X_1 = 0) f_{X_2}(X_2 = 1) = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8} \neq \frac{1}{4} \Rightarrow$ Not Independent

Q3: See page 63 bottom

Q2 Solution.

$PVE_j = \frac{\|\ell_j\|^2}{tr(\Sigma)}$ ← (jth Factor Explained Variance.) $\left(\ell_j^T \ell_j = \|\ell_j\|^2\right)$

For $L=(\ell_1 \dots \ell_r)$ $\quad \sum_{j=1}^{p} \lambda_j$ = total variance

Note that: In FA via PCA, $L = (\sqrt{\lambda_1} v_1 \dots \sqrt{\lambda_r} v_r)$, So the j^{th} load vector is defined by $\sqrt{\lambda_j}\vec{v_j}$, For λ_j to be jth biggest eigenvalue. v_j is the normalized (Unit) Eigenvector, i.e. $v_j^T v_j = 1$

So $\|\ell_j\|^2 = \ell_j^T \ell_j = (\sqrt{\lambda_j} v_j)^T (\sqrt{\lambda_j} v_j)$

$= (\sqrt{\lambda_j})^2 (v_j^T v_j) = (\sqrt{\lambda_j})^2 = \lambda_j$ ($v_j^T v_j = 1$)

$\Rightarrow PVE_j = \frac{\lambda_j}{tr(\Sigma)}$,

Which means PVE_j is $\lambda_j / tr(\Sigma)$ for jth eigenvalue of Σ.

Q3 Solution

i.e. Never change

Q: Is PVE overall and PVE_j Vs. all invariant after Rotation?

A: Yes. For overall, No for PVE_j.

Proof. WTS. $\|\tilde{L}\|_F^2 = \|L\|_F^2$. By definition, $\|\tilde{L}\|_F^2 = {}^{trace}(\tilde{L}\tilde{L}^T)$

① $trace(\tilde{L}\tilde{L}^T) = trace(LQ)(LQ)^T = {}^{tr}(LQQ^TL^T) = {}^{tr}(LL^T) = \|L\|_F^2$

(Note: orthogonal Q, $QQ^T = Q^TQ = I$) $\rightarrow = I$ For orthogonal Q

② WTS $\|\tilde{\gamma}_j\|_F^2 \neq \|L_j\|_F^2$.

For $\tilde{L} = (\tilde{\ell}_1 \dots \tilde{\ell}_r) = (Lq_1, \dots, Lq_r)$ (q_r: r^{th} column of Q)

$L \in \mathbb{R}^{p \times r}$, $L \times q_j \in \mathbb{R}^{p \times 1}$ matrix

For $\|\tilde{\ell}_j\|^2$,

$= \tilde{\ell}_j^T \tilde{\ell}_j = (Lq_j)^T (Lq_j) = q_j^T L^T L q_j$, this expression can take different value depending on q_j

Q4. Identifiability:

In factoranalysis model, if given Σ and find solution of L and Ψ, then we are solving a system of equations.

OFF-DIAGONAL Elements are decided by LL^T, while diagonal elements are decided by both Ψ and LL^T For Ψ is a diagonal matrix. $Var(X_i) = trace(\Sigma) = tr(LL^T) + tr(\Psi)$

$= \sum \ell_{ik}^2 + \Psi_i$

Q4

Consider when $r=1$, r is dimensionality of latentvector Z. So $r=1, p=3 \Rightarrow L = \begin{pmatrix} \ell_{11} \\ \ell_{21} \\ \ell_{31} \end{pmatrix} \in \mathbb{R}^{3\times 1}$ (ℓ_{31}: data points)

In this case, $LL^T = \begin{pmatrix} \ell_{11} \\ \ell_{21} \\ \ell_{31} \end{pmatrix} (\ell_{11}\ \ell_{21}\ \ell_{31}) = \begin{pmatrix} \ell_{11}^2+\Psi_1 & \ell_{11}\ell_{21} & \ell_{11}\ell_{31} \\ \ell_{21}\ell_{11} & \ell_{21}^2+\Psi_2 & \ell_{21}\ell_{31} \\ \ell_{31}\ell_{11} & \ell_{31}\ell_{21} & \ell_{31}^2+\Psi_3 \end{pmatrix} \in \mathbb{R}^{3\times 3}$

For $r=1$, the correlation between, for example, Σ_{11} and 2, will be explained solely by Σ_{12}, which is $\ell_{11}\ell_{21}$.

$\Sigma = LL^T + \Psi \Leftrightarrow \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{pmatrix}$ ⇒ Solve non-diagonal elements First

For $\begin{pmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{pmatrix} \Rightarrow \ell_{11}^2 = \frac{(\ell_{11}\ell_{21})(\ell_{11}\ell_{31})}{\ell_{21}\ell_{31}} = 1.575 \Rightarrow \ell_{11} \approx 1.255$ (存在解)

However, $Cov(X_1, Z_1) = E(X_1 Z_1) = E[(LZ+\varepsilon)Z_1] = \ell_{11} E[Z_1^2] = \ell_{11}$

And $Cov(X_1, Z_1) = \frac{Cov(X_1, Z_1)}{\sqrt{VarX_1}\sqrt{VarZ_1}} = \frac{\ell_{11}}{1 \cdot 1} = \ell_{11}$, but

$\ell_{11} > 1 \approx 1.255$, and

We conclude a contrid

and thus $\ell_{11} > 1$ i

not possible.

2 $\Sigma_{11} = \ell_{11}^2 + \Psi_1$

$1 = 1.575 +$

$\Rightarrow \Psi_1 < 0$

but Var(ε

As variance

non-negati

Conclude a

Q5 SCORING.

The joint Distribution is: $\begin{pmatrix} X \\ Z \end{pmatrix} \sim N_{p+r}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} LL^T+\Psi & L \\ L^T & I \end{pmatrix}\right)$

For $Z|X$, we have

$$\begin{cases} \mu_{Z|X} = \mu_Z + \Sigma_{ZX}\Sigma_{XX}^{-1}(X-\mu_X) \\ \Sigma_{Z|X} = \Sigma_{ZZ} - \Sigma_{ZX}\Sigma_{XX}^{-1}\Sigma_{XZ} \end{cases}$$

The score:

$$E(Z|X) = 0 + L^T(LL^T+\Psi)^{-1}\cdot(X-0) = L^T(LL^T+\Psi)^{-1}X$$

$$V(Z|X) = I - L^T(LL^T+\Psi)^{-1}L$$

We seek for $\hat{Z} = E(Z|X)$ s.t. $E\|Z-\hat{Z}\|^2$ is minimized (具体数学距期待补充)

(Regression

Scoring via Bayesian Stats)