Ch 1

Notation: P.1 is from class notes.

((P.3)) is from Johnson and Wichern 1988 (JW).

1] P.1 Multivariate analysis is used to analyze possibly correlated data containing $p \geq 2$ variables measured on a set of n cases.

2] *p1,10 A case or observation consists of p random variables measured for one person or thing. The ith case

$\underline{x}_i = (x_{i1}, \ldots, x_{ip})^T$. $\quad$ $\underline{x}_i = \underline{x}_i, ex)$ after (3)

Notation: Usually lowercase bold face letters $\underline{x}$ are column vectors while uppercase bold face letters $\underline{A}$ denote matrices with 2 or more columns. 2 exceptions:

i) $\underbrace{\underline{Y} \mid \underline{X} = \underline{x}}_{\text{Vectors}}$

ii) A multivariate location and dispersion estimator is $(\underset{\uparrow}{T}, C)$ $\quad$ T for test statistic

column vector, not boldface

3] p 2 The $n \times p$ <u>data matrix</u>

$$
W = \begin{pmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{pmatrix} = (\underline{v}_1 \cdots \underline{v}_p) = \begin{array}{c} \text{variable} \\ \begin{array}{ccc} 1 & & p \end{array} \\ \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \end{array} \begin{array}{c} \text{case} \\ 1 \\ \\ n \end{array}
$$

(JW uses $X = W^T$.)

		height	weight	age	
ex]	chris	69	135	22	case 1
	⋮				
	cindy	70	130	19	case n

$\underbrace{\qquad\qquad\qquad}_{\text{3 variables,}}$ expect height and weight to be correlated

4) * p 2 <u>Robust multivariate analysis</u> consists of techniques that are robust to

i) nonnormality

ii) outliers (these techniques are also often robust to nonnormality)

Read § 1.2 carefully (things that can go wrong)

§ 1.3 has some optimization results that will be used later.

5] p. 4: $p = 1$ gives the <u>location model</u> $Y_i = \mu + e_i$, $i = 1, \ldots, n$

6) P 5-6 Know: Univariate data $Y_1, \ldots, Y_n$

Sample mean $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$

order statistics $Y_{(1)}, \ldots, Y_{(n)}$ arrange data in ascending order $Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(n)}$

Sample median $MED(n) = Y_{\left(\frac{n+1}{2}\right)}$, n odd

middle ordered value $\dfrac{Y_{(\frac{n}{2})} + Y_{(\frac{n}{2}+1)}}{2}$, n even

Sample median absolute deviation

$$MAD(n) = MED\left(|Y_i - MED(n)|, i = 1, \ldots, n\right).$$

Sample variance $S^2 = \dfrac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1} = \dfrac{\sum Y_i^2 - n(\bar{Y})^2}{n-1}$

(like $EY^2 - (EY)^2$)

Sample standard deviation $S = \sqrt{S^2}$

ex] dataset 66, 3, 8, 5, 2

Find a) $\bar{Y}$, b) S, c) $MED(n)$ d) $MAD(n)$

usually omit

a) $\bar{Y} = \dfrac{\sum Y_i}{n} = \dfrac{66+3+8+5+2}{5} = \dfrac{84}{5} = \boxed{16.8}$

b) $S^2 = \dfrac{\sum Y_i^2 - n(\bar{Y})^2}{n-1} = \dfrac{4458 - 5(16.8)^2}{4} = \dfrac{3046.8}{4} = 761.7$

so $S = \sqrt{761.7} = \boxed{27.5989}$

c) sort data 2, 3, 5, 8, 66 $\quad MED(n) = \boxed{5}$

d) $Y_i - MED(n)$: $-3, -2, 0, 3, 61$ 2.5

sort $|Y_i - MED(n)|$: $0, 2, 3, 3, 61$

$$\boxed{MAD(n) = 3}$$

Skip § 1.5 ← section

ch 2

1] * p. 10 An important <u>multivariate location and dispersion</u> model uses a joint pdf $f(\underset{\sim}{z} \mid \underset{\sim}{\mu}, \Sigma)$ for a $p \times 1$ random vector $\underset{\sim}{x}$ that is completely specified by a $p \times 1$ population <u>location vector</u> $\underset{\sim}{\mu}$ and a $p \times p$ symmetric positive definite <u>dispersion matrix</u> Σ.

ex] $X_1, \ldots, X_p$ are iid (independent, identically distributed) with pdf $f(\underset{\sim}{z})$: Then $f(\underset{\sim}{z}) = \prod_{i=1}^{p} f(z_i)$. In this class, $\underset{\sim}{x} = (x_1, \ldots, x_p)^T$ where $x_1, \ldots, x_p$ are usually correlated, not independent.

2] Know for final p. 11-12 (54-55)

If second moments exist, the population mean (location)

$$E(\underline{x}) = \underset{\text{MU}}{\underset{\sim}{\mu}} = E\begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} E(x_1) \\ \vdots \\ E(x_p) \end{pmatrix}$$

and the $p \times p$ population covariance matrix (dispersion)

$$\text{Cov}(\underline{x}) = \overset{\text{sigma}}{\Sigma_{\underline{x}}} = E[(\underset{\sim}{x} - E(\underline{x}))(\underline{x} - E(\underline{x}))^T]$$

$$= E(\underline{x}\,\underline{x}^T) - E(\underline{x})\,E(\underline{x}^T) = (\sigma_{ij}).$$

So the ij entry of $\text{Cov}(\underline{x})$ is $\text{cov}(x_i, x_j) = \sigma_{ij}$.

The $p \times p$ population correlation matrix (dispersion)

$$\text{Cor}(\underline{x}) = \overset{\text{rho}}{\rho_{\underline{x}}} = (\rho_{ij}).$$ So the ij entry of $\text{Cor}(\underline{x})$ is $\rho_{ij} = \text{cor}(x_i, x_j)$.

3] p12 Let $z_i = \dfrac{x_i - E(x_i)}{\sqrt{\sigma_{ii}}}$ and $\underset{\sim}{z} = (z_1, \ldots, z_p)^T$.

Then $\text{Cov}(\underset{\sim}{z}) = \text{cor}(\underline{x}) = \rho_{\underline{x}}$.

4) * Let $\underline{x}$ be $p \times 1$ and $\underline{y}$ $q \times 1$. Assuming expected values exist, the $p \times q$ pop. covariance matrix of $\underline{x}$ and $\underline{y}$ is

$$\text{Cov}(\underline{x}, \underline{y}) = E(\underline{x} - E(\underline{x}))(\underline{y} - E(\underline{y}))^T = E\,\underline{x}\,\underline{y}^T - E(\underline{x})\,E(\underline{y}^T) = \Sigma_{\underline{x}\underline{y}}$$

5) p13 Let $x_{1j}, \ldots, x_{nj}$ be measurements on the jth random variable, $j = 1, \ldots, p$.

The jth sample mean $= \frac{1}{n} \sum_{k=1}^{n} x_{kj} = \overline{x}_j$.

The sample covariance $S_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)$ estimates σ_{ij}.

$S_{ii} = S_i^2$ is the sample variance that estimates $\sigma_{ii} = \sigma_i^2$. The sample correlation

$$r_{ij} = \frac{S_{ij}}{S_i \, S_j} = \frac{S_{ij}}{\sqrt{S_{ii} \, S_{jj}}}.$$

6) Know for final p 13-14

The sample mean $\overline{\underline{x}} = \frac{1}{n} \sum_{i=1}^{n} \underline{x}_i = \begin{pmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_p \end{pmatrix}$.

The sample covariance matrix $\underline{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})^T$

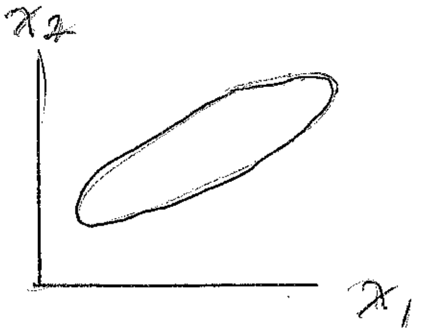
$= (S_{ij})$.

The sample correlation matrix $\underline{R} = (r_{ij})$.

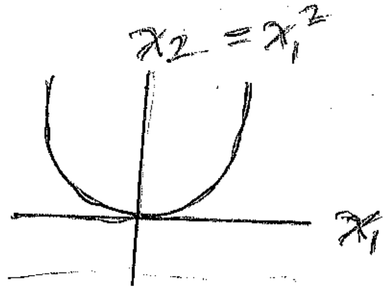
7) * p13 $z_i = \frac{x_i - \overline{x}_i}{\sqrt{S_{ii}}}$ are the standardized random variables.

8) The Correlation $cor(x_i, x_j)$ or r_{ij} measures the strength of the linear relationship between x_i and x_j.

ex)



$cor(x_1, x_2)$ is strong and positive



Strong nonlinear association but $cor(x_1, x_2) = 0$ is possible

9) <u>Know for final</u> p14-15

pop squared Mahalanobis distance

$$D^2_{\underline{x}}(\underline{\mu}, \Sigma) = (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu})$$

10) p.15 $A\underline{x} = \lambda \underline{x}$ if $(\lambda, \underline{x})$ is an eigenvalue eigenvector pair.

11) A ($p \times p$ symmetric) has p real eigenvalues and eigenvectors $\underset{\sim}{e}_i$ are chosen to be orthogonal (orthonormal), so $\underset{\sim}{e}_i^T \underline{e}_j = 0$ $i \neq j$, $\underline{e}_i^T \underline{e}_i = 1$.

If A is <u>positive semidefinite</u>, take

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

" " > 0 for A <u>positive definite</u>.

12] P15 (50–51) $\underset{\sim}{A}$ $p \times p$ symmetric $(\lambda_1, \underset{\sim}{e}_1), \ldots, (\lambda_p, \underset{\sim}{e}_p)$

Spectral decomposition of A is

$$\underset{\sim}{A} = \sum_{i=1}^{p} \lambda_i \underset{\sim}{e}_i \underset{\sim}{e}_i^T = \underset{\sim}{P} \underset{\sim}{\Lambda} \underset{\sim}{P}^T$$

where $\underset{\sim}{P} = [\underset{\sim}{e}_1 \; \underset{\sim}{e}_2 \; \ldots \; \underset{\sim}{e}_p]$ is a $p \times p$ orthogonal matrix and $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$.

13] * p15 The square root matrix of $p \times p$ positive definite symmetric matrix $\underset{\sim}{A}$ is $\underset{\sim}{A}^{\frac{1}{2}} = \underset{\sim}{P} \Lambda^{\frac{1}{2}} \underset{\sim}{P}^T$ where

$\Lambda^{\frac{1}{2}} = diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p})$.

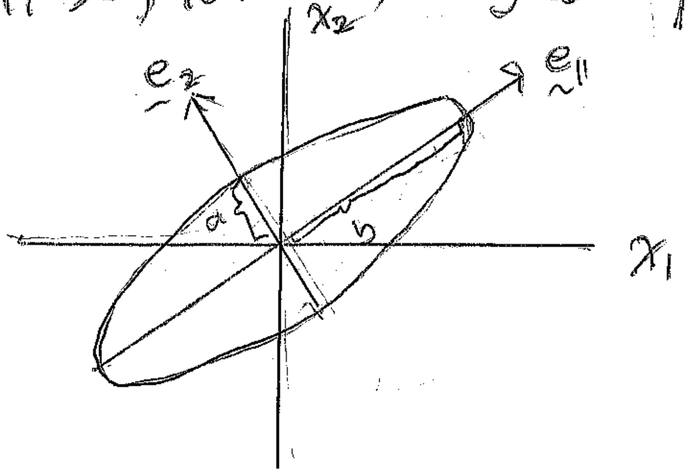
14] P16–17 * Let $U = D_{\underset{\sim}{x}}^2(\underset{\sim}{\mu}, \Sigma) = (\underset{\sim}{x} - \underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu})$

(Σ could be any positive definite $p \times p$ matrix eg a dispersion matrix).

Then $\{\underset{\sim}{x} : D_{\underset{\sim}{x}}^2 \le h^2\}$ is a hyperellipsoid centered at $\underset{\sim}{\mu}$. If $\underset{\sim}{\mu} = \underset{\sim}{0}$, then the axes of the hyperellipsoid are in the direction of $\pm \underset{\sim}{e}_i$ with half length $= h\sqrt{\lambda_i}$.

The volume of the hyperellipsoid $= \dfrac{2\pi^{p/2}}{p\Gamma(\frac{p}{2})} |\Sigma|^{\frac{1}{2}} h^p$.

see (49-50, 102-103) JW p50



$a = h\sqrt{\lambda_2}$

$b = h\sqrt{\lambda_1}$

If $h^2 = U_{1-\alpha}$ where $P(U \le u_{1-\alpha}) = 1-\alpha$

and the $\underset{\sim}{x} \sim \underbrace{EC_p(\underset{\sim}{\mu}, \Sigma, g)}_{\text{see Def 3.3}}$ where g

is continuous and decreasing, then $\{\underset{\sim}{x} \mid D_{\underset{\sim}{x}}^2 \le U_{1-\alpha}\}$

is the pop highest density $100(1-\alpha)\%$ region.

So the data cloud looks like a hyperellipsoid where the axes directions and lengths depend on the eigenvalue eigenvector pairs $(\lambda_1, \underset{\sim}{e}_1), \ldots, (\lambda_p, \underset{\sim}{e}_p)$ of the dispersion matrix Σ; this information tells you how the data disperses in $\mathbb{R}^p$.

15] p117 $\{\underset{\sim}{x} : D_{\underset{\sim}{x}}^2(\bar{\underset{\sim}{x}}, S) \le h^2\}$ is a hyperellipsoid centered at $\bar{\underset{\sim}{x}}$ with axes directions parallel to $\hat{\underset{\sim}{e}}_i$, half lengths $h\sqrt{\hat{\lambda}_i}$ where $(\hat{\lambda}_1, \hat{\underset{\sim}{e}}_1), \ldots, (\hat{\lambda}_p, \hat{\underset{\sim}{e}}_p)$ are the eigenvalue eigenvector pairs of S.

Vol of hyperellipsoid $= \frac{2\pi^{p/2}}{p\,\Gamma(p/2)} |S|^{\frac{1}{2}} h^p$.

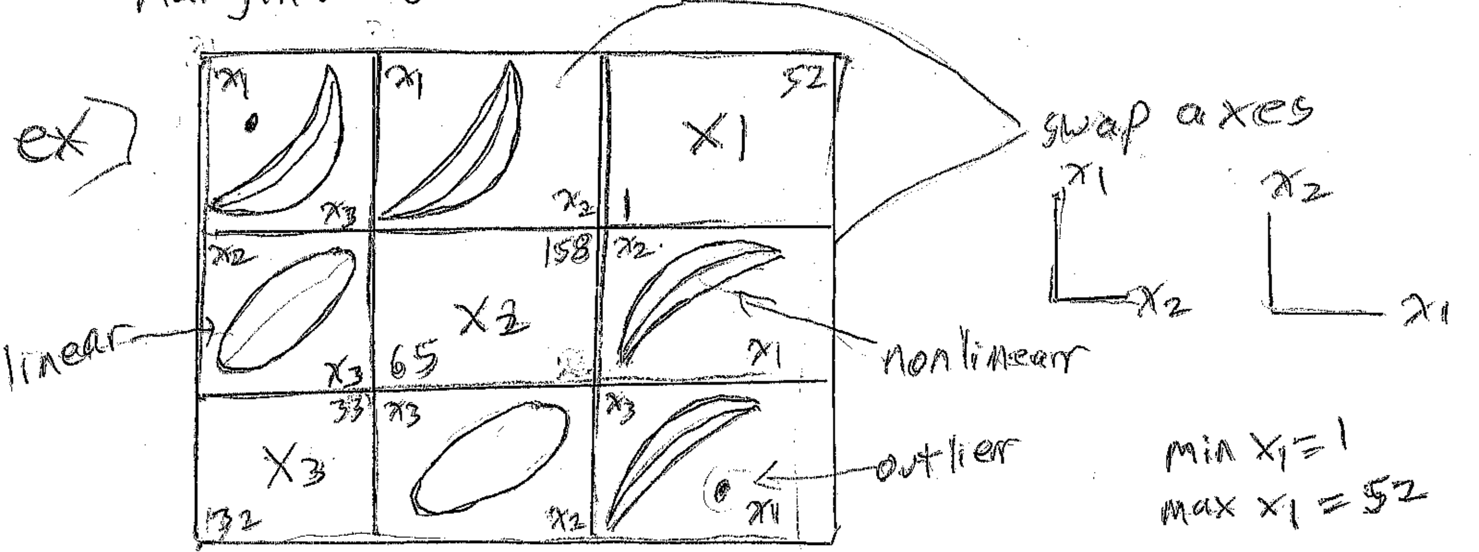
16] p17 (103-106) $|S| = \det(S)$ = generalized sample variance (5.5

§2.4 17] p18 Correlation and covariance matrices make the most sence when X_i and X_j are linearly related for $i \neq j = 1, \ldots, p$. Transformations are used to remove gross nonlinearities. These methods are also useful for regression models that use p continuous predictors $\underline{x} = (x_1, \ldots, x_p)^T$.

18] p18 A scatterplot of x vs Y is used to visualize $Y|x$.

19] p18 Know A scatterplot matrix is an array of scatterplots used to examine marginal bivariate relationships between predictors.

ex]



min $X_1 = 1$
max $X_1 = 52$

20} P18 <u>know</u> <u>power transformation</u>

Assume all values of $x > 0$.

$x = t_\lambda(w) = w^\lambda, \quad \lambda \neq 0$

$x = t_0(w) = \log(w), \quad \lambda = 0$

21} P18 The <u>ladder of powers</u> $\Lambda_L = \{-1, -\frac{1}{2}, -\frac{1}{3}, 0, \frac{1}{3}, \frac{1}{2}, 1\}$.

Can add powers like $\pm 2, \pm 3$ if necessary.

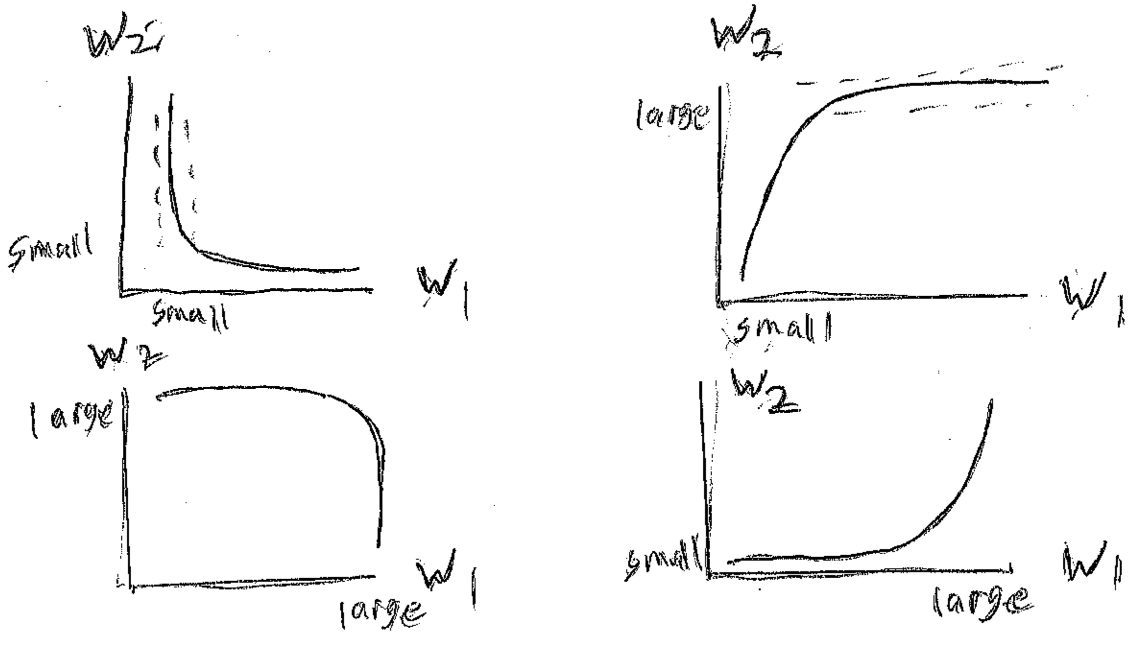
Often use power transformations to linearize the scatterplot matrix.

For 22) – 25) plot of w_1 vs w_2 where $w_1 > 0$ $w_2 > 0$.

22} <u>know for final</u> p21 <u>ladder rule</u>: In a plot of W_1 vs W_2 where $W_1 > 0$ and $W_2 > 0$, linearize with a plot of $W_1^{\lambda_1}$ vs $W_2^{\lambda_2}$ (or $\log(w)$).

To spread smaller values of w_i, make λ_i smaller.

larger larger.



project cloud on axis, narrow slices with lots of data need spreading

6.5

23] Know for final p20 log rule:

Suppose $W > 0$. If $\frac{\max(W)}{\min(W)} > 10$, use

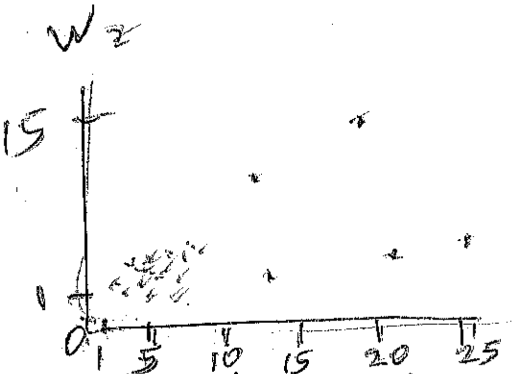
$X = \log(W)$

24] Unit rule: if W_1 and W_2 have the same units, use the same transformation of W_1 and W_2.

25] range rule: if $\frac{\max(W)}{\min(W)} < 2$, keep W ($X = W$). no transformation

26] Sometimes several power transformations work. (subject matter experts) Lacking theory, no transformation $X = W$ is best, then $X = \log(W)$, $X = \sqrt{W}$ and $X = \frac{1}{W}$.

ex]



both W_1 and W_2 are right skewed

small values of W_1 and W_2 need spreading

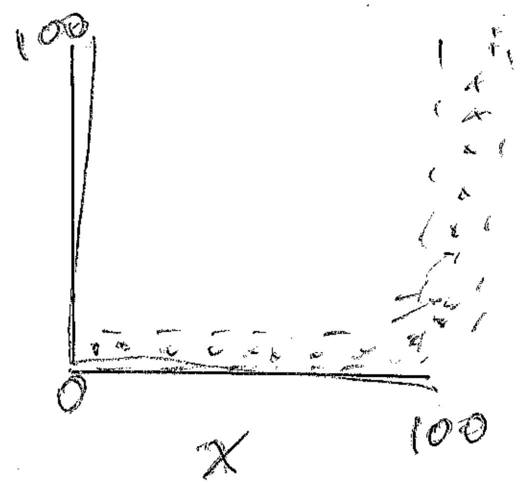
$\frac{\max W_1}{\min W_1} \approx \frac{25}{1} > 10$

$\frac{\max W_2}{\min W_2} \approx \frac{15}{1} > 10$

so $\log(W_1) = X_1$ vs $\log(W_2) = X_2$ might work ex] on exam problem, there may be a choice between 2 transformations, and a plot. log rule or ladder rule is used to determine which transformation is better.

MV7

ex]



In the plot of x vs y above, $\lambda=1$.
which transformation will increase the linearity of the plot, $\log(x)$ or x^2?
Explain. $\lambda=0$ $\lambda=2$

soln] $\boxed{x^2}$, want to spread large values of x
so make λ larger ($\log x \to \lambda=0$, $x^2 \to \lambda=2$).

Ch 3

Notation for ch 3: $\underset{\sim}{X}$, $\underset{\sim}{Y}$ for random vectors

so $\underset{\sim}{Y} \mid \underset{\sim}{X} = \underset{\sim}{x}$ makes sense.

1] Know p30 Let $\underset{\sim}{x}$ and $\underset{\sim}{y}$ be $p\times 1$ random vectors, $\underset{\sim}{a}$ a conformable constant vector, and let A and B be conformable constant matrices. Then

$$E(\underset{\sim}{x}+\underset{\sim}{y}) = E(\underset{\sim}{x}) + E(\underset{\sim}{y})$$

$$E(\underset{\sim}{a}+\underset{\sim}{x}) = \underset{\sim}{a} + E(\underset{\sim}{x})$$

$$E(A\,\underset{\sim}{x}\,B) = A\,E(\underset{\sim}{x})\,B$$

$$\mathrm{cov}(\underset{\sim}{a}+A\underset{\sim}{x}) = \mathrm{cov}(A\underset{\sim}{x}) = \underset{q\times p}{A}\ \mathrm{cov}(\underset{\sim}{x})\ \underset{p\times q}{A^T}$$

Useful for HW1B

2] p29 A $p\times 1$ random vector $\underset{\sim}{x}$ has a p-dimensional multivariate normal (MVN) distribution, $\underset{\sim}{x} \sim N_p(\underset{\sim}{\mu}, \Sigma)$, iff $\underset{\sim}{t}^T\underset{\sim}{x}$ has a univariate normal distribution for any $p\times 1$ constant vector $\underset{\sim}{t}$. $E(\underset{\sim}{x}) = \underset{\sim}{\mu}$ and $\mathrm{Cov}(\underset{\sim}{x}) = \Sigma_{\underset{\sim}{x}} = \Sigma$.

3]. Usually want Σ positive definite so $\underset{\sim}{x}$ has pdf

$$f(\underset{\sim}{z}) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\underset{\sim}{z}-\underset{\sim}{\mu})^T\Sigma^{-1}(\underset{\sim}{z}-\underset{\sim}{\mu})\right]$$ where $|\Sigma| = \det(\Sigma)$.