28] p52 Continuous Mapping Th) If $X_n \xrightarrow{D} X$ and g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

29] p51 $X_n \xrightarrow{P} \tau(\theta)$ iff $X_n \xrightarrow{D} \tau(\theta)$

30] * p48 W_n is bounded in probability, $W_n = O_p(1)$, if $\forall \varepsilon > 0 \ \exists \ D_\varepsilon, N_\varepsilon > 0 \ni$

$P(|W_n| \le D_\varepsilon) \ge 1-\varepsilon \quad \forall n \ge N_\varepsilon$.

$W_n = O_p(X_n)$ if $\left|\frac{W_n}{X_n}\right| = O_p(1)$ big O_p

where $X_n = n^{-\delta}$ is common.

little o_p

31] * p49 $W_n = o_p(n^{-\delta})$ if $n^\delta W_n = o_p(1)$ which means $n^\delta W_n \xrightarrow{P} 0$.

32] * p49 W_n has the same order as X_n in probability

$W_n \asymp_p X_n$, if $\Big(\forall \varepsilon > 0 \ \exists \ N_\varepsilon > 0$ and $0 < d_\varepsilon < D_\varepsilon \ni$

$P\left(d_\varepsilon \le \left|\frac{W_n}{X_n}\right| \le D_\varepsilon\right) = P\left(\frac{1}{D_\varepsilon} \le \left|\frac{X_n}{W_n}\right| \le \frac{1}{d_\varepsilon}\right) \ge 1-\varepsilon\Big)$

$\forall n \ge N_\varepsilon$. Or) $W_n = O_p(X_n)$ and $X_n = O_p(W_n)$.

33] If $W_n \asymp_p n^{-\delta}$ for $\delta > 0$, then W_n has rate n^δ.

34] If $n^\delta (W_n - \tau) \xrightarrow{D} X$ for some nondegenerate RV X, then W_n has rate n^δ. $n^{1/2}$ rate $= n^{1/2}$ consistency is good.

35] p54 Multivariate limit theorems (12.5)

Let $\underset{\sim}{X}_n$ have joint cdf $F_n(\underset{\sim}{x})$ and $\underset{\sim}{X}$ have joint cdf $F(\underset{\sim}{x})$.

a) $\underset{\sim}{X}_n \xrightarrow{D} \underset{\sim}{X}$ if $F_n(\underset{\sim}{x}) \to F(\underset{\sim}{x})$ at all points $\underset{\sim}{x}$ where $F(\underset{\sim}{x})$ is continuous.

b) $\underset{\sim}{X}_n \xrightarrow{P} \underset{\sim}{X}$ if $\forall \varepsilon > 0$, $P(\|\underset{\sim}{X}_n - \underset{\sim}{X}\| > \varepsilon) \to 0$ as $n \to \infty$.

36] Know Multivariate CLT (MCLT)

If $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ are iid $k \times 1$ random vectors with $E(\underset{\sim}{X}) = \underset{\sim}{\mu}$ and $Cov(\underset{\sim}{X}) = \Sigma_{\underset{\sim}{X}}$, then

$$\sqrt{n}\,(\overline{\underset{\sim}{X}}_n - \underset{\sim}{\mu}) \xrightarrow{D} N_k(\underset{\sim}{0}, \Sigma_{\underset{\sim}{X}}).$$

37] p55 If estimator $\underset{\sim}{g}(\underset{\sim}{T}_n) \xrightarrow{P} \underset{\sim}{g}(\underset{\sim}{\theta})$ $\forall \underset{\sim}{\theta} \in \Theta$, then $\underset{\sim}{g}(\underset{\sim}{T}_n)$ is a consistent estimator of $\underset{\sim}{g}(\underset{\sim}{\theta})$.

38] If $n^{\delta}(\underset{\sim}{g}(\underset{\sim}{T}_n) - \underset{\sim}{g}(\underset{\sim}{\theta})) \xrightarrow{D} \underset{\sim}{X}$, then $\underset{\sim}{g}(\underset{\sim}{T}_n) \xrightarrow{P} \underset{\sim}{g}(\underset{\sim}{\theta})$.

39] If $\underset{\sim}{X}_n \xrightarrow{P} \underset{\sim}{X}$, then $\underset{\sim}{X}_n \xrightarrow{D} \underset{\sim}{X}$.

$\underset{\sim}{X}_n \xrightarrow{P} \underset{\sim}{g}(\underset{\sim}{\theta})$ iff $\underset{\sim}{X}_n \xrightarrow{D} \underset{\sim}{g}(\underset{\sim}{\theta})$.

40] p56 Let $\underset{r\times c}{A_n} = [a_{ij}(n)]$.

a) $A_n = O_p(X_n)$ if $a_{ij}(n) = O_p(X_n)$ for $1 \le i \le r$, $1 \le j \le c$

b) $\mathcal{O}_p(X_n)$ $\mathcal{O}_p(X_n)$

c) $A_n \asymp_p \frac{1}{g(n)}$ $a_{ij}(n) \asymp_p \frac{1}{g(n)}$

d) If $\underbrace{T_n - \underset{\sim}{\mu}}_{\text{vectors}} \asymp_p \frac{1}{g(n)}$ and $C_n - c\Sigma \asymp_p \frac{1}{g(n)}$ where $c>0$,

then (T_n, C_n) has rate $g(n)$ and $g(n) = \sqrt{n}$ is good.

41] p57 continuous mapping th: If $\underline{X}_n \xrightarrow{D} \underline{X}$

then $\underset{\sim}{g}(\underline{X}_n) \xrightarrow{D} \underset{\sim}{g}(\underline{X})$ for continuous fn $\underset{\sim}{g}$.

42] Know a) If $\sqrt{n}\,(T_n - \underset{\sim}{\mu}) \xrightarrow{D} N_p(\underline{\theta}, \Sigma)$ and

A is a $q\times p$ constant matrix, then

$A\sqrt{n}(T_n - \underset{\sim}{\mu}) = \sqrt{n}\,(A\,T_n - A\underset{\sim}{\mu}) \xrightarrow{D} A[N_p(\underline{\theta}, \Sigma)]$

$\sim N_q(A\underline{\theta}, A\Sigma A^T)$.

MVN and $\xrightarrow{D}$ MVN are similar

b) Let $\Sigma > 0$. If $\sqrt{n}\,(T_n - \underset{\sim}{\mu}) \xrightarrow{D} N_p(\underset{\sim}{0}, \Sigma)$ and if

C is a consistent estimator of Σ, then

$n\,(T - \underset{\sim}{\mu})^T C^{-1} (T - \underset{\sim}{\mu}) \xrightarrow{D} \chi^2_p$.

43) * If $\underline{X}_n \xrightarrow{D} \underline{X}$

then $A\underline{X}_n \xrightarrow{D} A\underline{X}$ for conformable constant matrix A by the continuous mapping theorem.

E1 Q3 { ex) see HW3A.
{ $\boxed{\text{ex}}$ Suppose $\underline{X}_1, \ldots, \underline{X}_n$ are iid

where $\underline{X}_i = \begin{pmatrix} X_{i1} \\ X_{i2} \end{pmatrix}$ and $X_{i1} \sim \text{Pois}(\lambda_1) \perp\!\!\!\perp X_{i2} \sim \text{Pois}(\lambda_2)$. Recall that if $W \sim \text{pois}(\lambda)$ then $E(W) = V(W) = \lambda$. Find the limiting distribution of $\sqrt{n}\,(\overline{\underline{X}} - \underline{c})$ for appropriate vector $\underline{c}$.

Soln: $\sqrt{n}\,(\overline{\underline{X}} - E\underline{X}) \xrightarrow{D} N_p(\underline{0}, \Sigma_{\underline{X}})$

So need enough info to get $E(\underline{X}) = \underline{\mu}$ and $\Sigma_{\underline{X}} = \text{cov}(\underline{X})$. Now $E(\underline{X}) = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \underline{c}$ and $\Sigma_{\underline{X}} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ since $X_{i1} \perp\!\!\!\perp X_{i2}$, so $\text{cov}(X_{i1}, X_{i2}) = 0$.

So $\sqrt{n}\left[\overline{\underline{X}} - \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}\right] \xrightarrow{D} N_2\left[\underline{0}, \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}\right]$

Ch 4 1) Know for Final, Q3

MV 14

$\underline{x}$ is the sample mean and

$T = MED(W)$ is the coordinatewise median.

To find $\overline{x}$ take the sample mean of each variable and collect in a vector. To find MED(W), find the median of each variable and collect in a vector.

ex) See problem 3.6 and HW3 B.

ex)

length x_1	nasal ht x_2	bigonal x_3	
168	51	104	← 1st case (person)
176	55	106	
179	55	104	$n = 5$
176	46	114	
177	48	101	5th case

$\sum x_{1j} = 876$ $\quad \sum x_{2j} = 255$ $\quad \sum x_{3j} = 529$

$$\overline{x} = \frac{1}{5}\begin{pmatrix} 876 \\ 255 \\ 529 \end{pmatrix} = \begin{pmatrix} 175.2 \\ 51.0 \\ 105.8 \end{pmatrix}$$

← sample mean of length

work

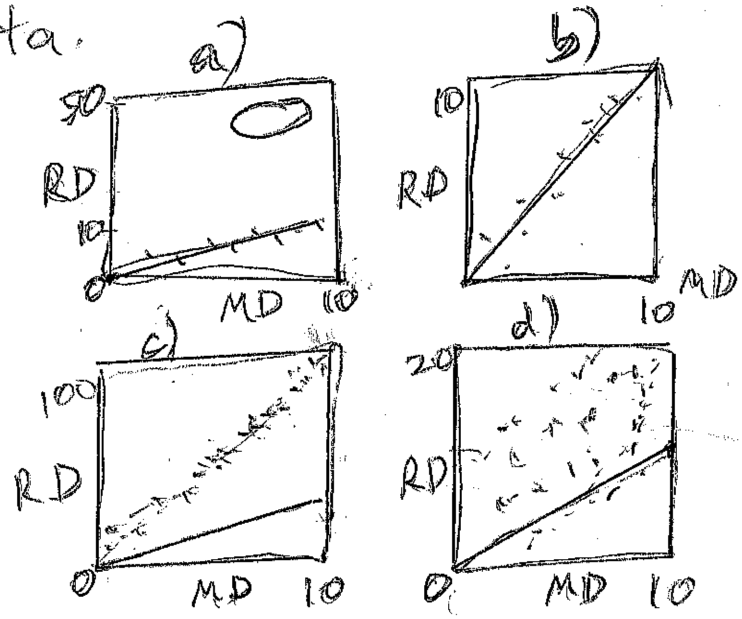
$$MED(W) = \begin{pmatrix} 176 \\ 51 \\ 104 \end{pmatrix}$$

168	176	176	177	179
46	48	51	55	55
101	104	104	106	114

2) Know for final P81, 110, 113 A DD plot is a plot of classical vs robust Mahalanobis distances with the identity line that has unit slope and zero intercept added as a visual aid. The DD plot is used to check i) if the data set is MVN (plotted points cluster about the identity line RD = MD). ii) If the data set is EC but not MVN (plotted points follow a line through the origin with slope > 1, or < 1 rare). iii) Data is not EC (Plotted points do not follow a line through the origin) iv) Outliers are present (eg som plotted points are far from the bulk of the data or the plotted points follow two lines).

3) * pvi **Outliers** are cases far from the bulk of the data.

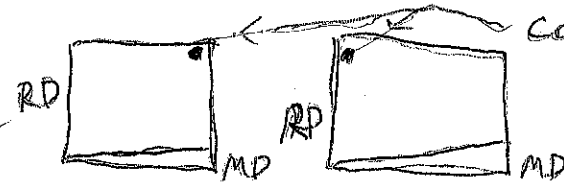
Exam type ex
see P111



what do the 4 DD plots show?

a) outliers b) MVN c) EC not MVN d) not EC

plots fill space



could miss these outliers if you do not know that plots fill space

4) Let $\underset{\sim}{\mu}$ be a $p \times 1$ location vector and Σ a $p \times p$ symmetric positive definite dispersion matrix. $\frac{p(p+1)}{2}$ parameters. There are $\frac{p(p+3)}{2}$ unknown parameters to be estimated.

$\bar{\underline{Z}} = \underline{0}$

sample mean of standardized data

5) Know p 62 For $(\bar{\underline{x}}, S)$ or $(\bar{\underline{Z}} = \underline{0}, R)$, want $n > 10p$.

For FCH, RFCH, RMVN, want $n \geq 20p$.

6) p63 (T, C) is affine equivariant if

$Z = \begin{pmatrix} \underline{z}_1^T \\ \vdots \\ \underline{z}_n^T \end{pmatrix}$ $\quad \underline{z}_i = A\underline{x}_i + \underline{b}$

$T(Z) = T(WA^T + \underline{1}\underline{b}^T) = A\, T(W) + \underset{\sim}{b}$ and

$C(Z) = C(WA^T + \underline{1}\underline{b}^T) = A\, C(W) A^T$, where $A > 0$ is $p \times p$.

7) If (T, C) is affine equivariant, $D_i^2(Z) = D_i^2(W)$

§4.2 8) p64 Let $W = \begin{pmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{pmatrix}$ be the data matrix.

Let $W_d^n = \begin{pmatrix} \underline{w}_1^T \\ \vdots \\ \underline{w}_n^T \end{pmatrix}$ denote the $n \times p$ data matrix after d_n of the $\underline{x}_i$ have been replaced by arbitrarily bad contaminated cases. The breakdown value of T at W is $B(T, W) = \min\left\{\frac{d_n}{n} : \sup_{W_d^n} \|T(W_d^n)\| = \infty\right\}$

and the breakdown value of C =

$$\min\left\{\frac{d_n}{n} : \sup_{W_d^n} \max\left[\frac{1}{\lambda_p[C(W_d^n)]}, \lambda_1[C(W_d^n)]\right] = \infty\right\}.$$

Want smallest and largest eigenvalues bounded away from 0 and ∞.

9) If no more than p cases of the clean data (the x_i) lie on any $p-1$-dimensional hyperplane, then the clean cases are in general position. Then (T, C) is high breakdown (HB) if $\gamma_n = \frac{d_n}{n} \to 0.5$ and zero breakdown if $\gamma_n \to 0$ as $n \to \infty$ where γ_n is the breakdown value of (T, C). (15.5)

ex) $\gamma_n = \frac{1}{n}$ for $(\bar{x}, S)$ since 1 case can make $\bar{x}$ arbitrarily large.

10) p xi, 100 Estimators with $O(n^4)$ or higher complexity will not be used, actually complexities higher than $O[(n^3 + n^2 p + np^2 + p^3) \log(n)]$ take too long to compute.

For brand name robust estimators, Stahel Donoho can be computed for $p \approx 2$. MVE and MCD can be computed in a few hours for $n = 100$ and $p \approx 6$ with a branch and bound algorithm, but can only be simulated for $p \le k$, $k \approx 6$. For general p, MVE complexity is exponential and MCD complexity is higher than $O(n^{p^2/2})$.

11) Rule of thumb Brand name estimators from the Rousseeuw Yohai paradigm that have been shown to be consistent and HB take too long to compute and are replaced by Fake-brand name estimators that do not have any theory. So papers that claim to use S, MVE, MCD etc are using Fake S, Fake MVE and Fake MCD.

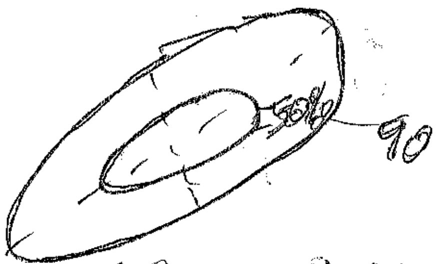
12] * See 35) – 40) on EI review

13] p68 $\det(S_{t+1,j}) \le \det(S_{t,j})$ with equality iff

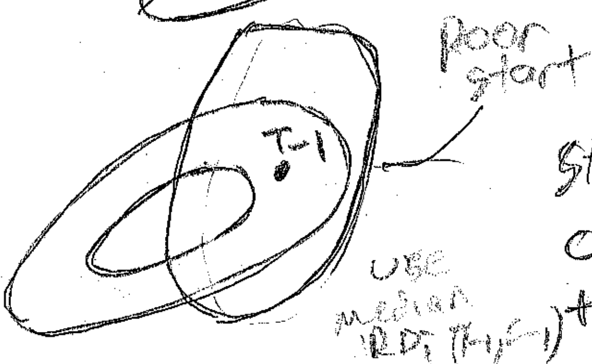
$(\bar{x}_{t,j}, S_{t,j}) = (\bar{x}_{t+1,j}, S_{t+1,j})$ for $t \ge 0$ where the jth start is $(T_{-1,j}, C_{-1,j})$.

So the determinant decreases until convergence for concentration.

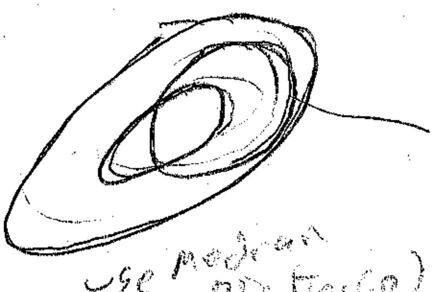
14]



For EC distributions (with $g \downarrow$), want to estimate $(\bar{x}_{50}, S_{50})$ applied to the highest 50% density region



step 0: scale ellipsoid to contain half the data. Compute the classical estimator $(\bar{x}_0, S_0)$ on the half set.



cover half the data

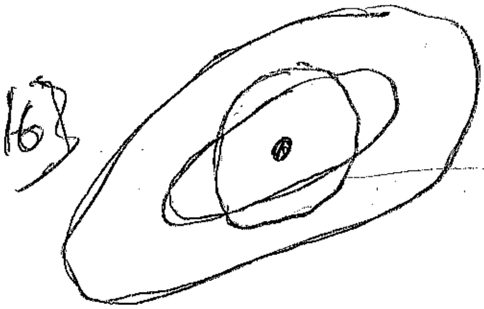
step 1 new half set is a lot better

get $(\bar{x}_1, S_1)$ from half set

at convergence, region for attractor seems pretty good.

15] DGK estimates HDR half set at step 1 for large n and a wide class of EC data. Start $(T_{-1}, C_{-1}) = (\bar{x}, S)$ = classical MLD estimator. DGK has fair outlier resistance.

Start $(T_{-1}, C_{-1}) = (MED(W), I_p)$

16] 

MB step 0 gets a hypersphere centered at the coordinatewise median that covers half the data. Get $(\bar{x}_0, S_0)$

can get even in high dimensions

step 1 ellipsoid covering $\frac{1}{2}$ the data based on $(\bar{x}_0, S_0)$ is better than the hypersphere but underestimates the major axis

Stop at step K so $(\bar{x}_K, S_K)$ is still biased but bias is small. The ellipsoid covering half the data is still underestimates the major axis determined by $(\hat{\lambda}_1, \hat{e}_1) \approx (\lambda_1, e_1)$ corresponding to $S \approx Cov(x) = \Sigma_x$.

The HBMB estimator is highly outlier resistant but biased for $(\mu, c\Sigma)$.

17] Σ determines the shape of the highest density region (hyperellipsoid) for EC distributions (with $g \downarrow$).

18] The (T_0, C_0) halfset may contain outliers, but after concentration the (T_K, C_K) halfset may not contain any outliers. See Hw 3 E, F and Fig 4.5 – 4.11.

19) FCH uses the DGK or MB estimator. FCH uses MB if $T_{DGK} > \text{MED } D_i(\text{MED}(W), I)$.

So T_{DGK} is further than half the data in Euclidean distance from MED(W). Otherwise FCH uses $\begin{cases} \text{DGK} & \text{if } \det(C_{DGK}) \le \det(MB) \\ \text{MB} & \quad > \end{cases}$

RFCH and RMVN are reweighted versions of FCH where RMVN estimates $(\underline{\mu}, \underline{\Sigma})$ for $N_p(\underline{\mu}, \underline{\Sigma})$ data even when certain types of outliers are present. FCH, RFCH and RMVN are $\sqrt{n}$ consistent estimators of $(\underline{\mu}, c_i \underline{\Sigma})$ on a large class of EC distributions where $c_i = 1$ for MVN data, $c_1 = $ FCH, $c_2 = $ RFCH, and $c_3 = $ RMVN.

20) "MCD estimator" in papers is actually the Fake-MCD estimator which has no theory.

<u>Stein Ch 4</u>

Ch 5) $\square^*$ DD plot of MD vs RD.

2) $MD_i = D_i(\overline{\underline{x}}, S)$, $RD_i = D_i(T, C)$

often (T, C) is FCH, RFCH, RMVN, scaled DGK, scaled MB or Fake MCD so plotted points follow identity line for MVN data.

<u>End Exam I material</u>

§5.2 3) p114 A large sample $100(1-\alpha)\%$ **prediction region** (PR) for a future observation (17.5)

$\underline{x}_f$ satisfies $1-\alpha_n = P(\underline{x}_f \in A_n) \to 1-\alpha$ as $n \to \infty$.

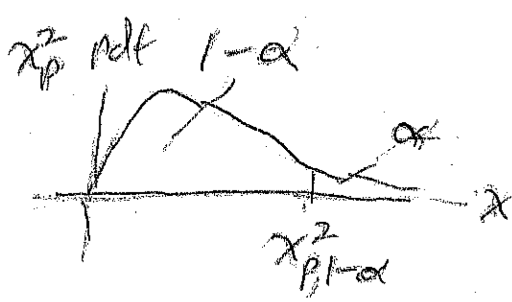
An asymptotically optimal $100(1-\alpha)\%$ prediction region has a volume converging to the volume of the population minimum volume (highest density) $100(1-\alpha)\%$ covering region.

4) p114 (p151) The classical parametric MVN prediction region is

$$\{\underline{z} \mid (\underline{z}-\underline{\bar{x}})^T S^{-1} (\underline{z}-\underline{\bar{x}}) \le \chi^2_{p,1-\alpha}\}$$

$$= \{\underline{z} \mid D^2_{\underline{z}}(\underline{\bar{x}}, S) \le \chi^2_{p,1-\alpha}\}$$ where

$$P(\chi^2_p \le \underbrace{\chi^2_{p,1-\alpha}}_{\text{JW notation: } \chi^2_p(\alpha)}) = 1-\alpha.$$



This region is asymptotically optimal if $\underline{x}_1, \ldots, \underline{x}_n, \underline{x}_f$ are iid $N_p(\underline{\mu}, \Sigma)$, but is not robust to nonnormality.

5) p114 (p126) The MVN highest density population region is $\{\underline{z} \mid D^2_{\underline{z}}(\underline{\mu}, \Sigma) \le \chi^2_{p,1-\alpha}\}$. See **HW2 A** for $p=2$, 10%, 30%, 50%, 70%, 90%, 95% regions.