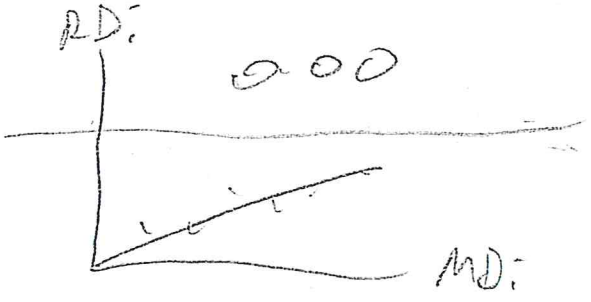
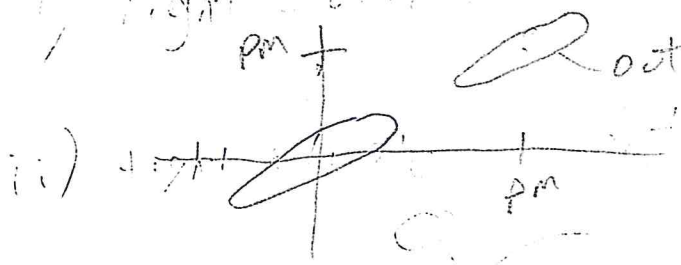
64) To compare outlier resistance of MB, FCH, RFCH, RMUN, DGK, MBA, FMCD etcetera, generate M outlier data sets and count how many times the RD; for all of the outliers > largest RD; of the clean data. Then the outliers can be separated from the clean data with a horizontal line in the DD plot.



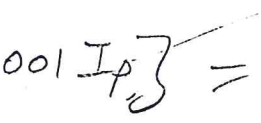
Let the data set have n cases and let γ = outlier percentage. Counts near M with γ near but less than 0.5 are good.

65) Clean cases: $\underset{\sim}{X} \sim N_p(\underset{\sim}{0}, \mathrm{diag}(1, \ldots, p))$.

a) mean shift outliers $\underset{\sim}{x} \sim N_p[\underset{\sim}{\mu_M}, \mathrm{diag}(1, \ldots, p)]$



b) $\underset{\sim}{X} \sim N[(0, \ldots, 0, \mu_M)^T, 0.001 I_p]$ = near point mass on major axis



If ... clean data is $N_p(\underline{0}, I_p)$ then outliers are like $\underline{x} \sim N_p\left((0,\ldots,0,\frac{pm}{\sqrt{p}})^T, dI_p\right)$

type (i) outliers

p	γ	n	pm	MBA	FCH	RFCH	RMVN	OGK	FMCD	MB
10	.1	100	40	73	92	92	92	100	95	100
10	.25	100	25	0	99	99	90	0	0	99
60	.4	200	150	0	100	100	100	0	0	100

MBA does not use location criterion and selects DGK too much! FCH uses location criterion and usually selects MB

see HWS: mldsim6

9/11/16 Large Sample Theory

1) Know Central Limit Theorem (CLT):
Let $Y_1, \ldots, Y_n$ be iid with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2$.
Then $\sqrt{n}(\bar{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2)$.

converges in distribution to

2) If $W_n \xrightarrow{D} X$, then X is the asymptotic distribution or limiting distribution of W_n.

X does not depend on n. The approximate distribution of $\bar{Y}$ is $\bar{Y} \sim AN(\mu, \frac{\sigma^2}{n}) \Rightarrow$

$\bar{Y} \simeq N(\mu, \frac{\sigma^2}{n})$, which does depend on n.

3) Let $\{Z_n, n = 1, 2, \ldots\}$ be a sequence of RVs with CDFs F_n and let X be a RV with CDF F. Then Z_n converges in dist to X, written $Z_n \xrightarrow{D} X$, if $\lim_{n \to \infty} F_n(t) = F(t)$ at each continuity point of F.

4) X_n converges in probability to X, written $X_n \xrightarrow{P} X$, if $\lim_{n \to \infty} P(|X_n - X| \geq \varepsilon) = 0 \quad \forall \varepsilon$

$\iff \lim_{n \to \infty} P(|X_n - X| < \varepsilon) = 1 \quad \forall \varepsilon > 0.$

5) A sequence of estimators T_n is a consistent

of $\tau(\theta)$ if $T_n \xrightarrow{P} \tau(\theta)$ $\forall \theta \in \Theta$

5) If $\lim_{n\to\infty} U_\theta(T_n) = 0$ and $\lim_{n\to\infty} E_\theta(T_n) = \tau(\theta)$ both $\forall \theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

Multivariate limit theorems

6) Let $\underline{X}_n$ have (joint) CDF $F_n(\underline{x})$ and $\underline{X}$ have (joint) CDF $F(\underline{x})$.

a) $\underline{X}_n \xrightarrow{D} \underline{X}$ if $F_n(\underline{x}) \to F(\underline{x})$ at all continuity points $\underline{x}$ of F.

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

7) Multivariate CLT (MCLT): If $\underline{X}_1, \ldots, \underline{X}_n$ are iid $k \times 1$ random vectors with $E(\underline{X}) = \underline{\mu}$ and $\mathrm{Cov}(\underline{X}) = \Sigma_{\underline{X}}$, then $\sqrt{n}(\bar{\underline{X}} - \underline{\mu}) \xrightarrow{D} N_k(\underline{\mu}, \Sigma_{\underline{X}})$.

8) a) If estimator $\underline{g}(\underline{T}_n) \xrightarrow{P} \underline{g}(\underline{\theta})$ $\forall \underline{\theta} \in \Theta$, then $\underline{g}(\underline{T}_n)$ is a consistent estimator of $\underline{g}(\underline{\theta})$.

b) If $n^\delta(\underline{g}(\underline{T}_n) - \underline{g}(\underline{\theta})) \xrightarrow{D} \underline{X}$, then $\underline{g}(\underline{T}_n) \xrightarrow{P} \underline{g}(\underline{\theta})$.

c) If $\underline{X}_n \xrightarrow{P} \underline{X}$ then $\underline{X}_n \xrightarrow{D} \underline{X}$.

d) $\underline{X}_n \xrightarrow{P} \underline{g}(\underline{\theta})$ iff $\underline{X}_n \xrightarrow{D} \underline{g}(\underline{\theta})$.

(10) Continuous Mapping Theorem.

Let g be a continuous function $g: \mathbb{R}^k \to \mathbb{R}^q$.

If $\underline{x}_n \xrightarrow{D} \underline{x}$, then $g(\underline{x}_n) \xrightarrow{D} g(\underline{x})$.

(11) (Know for exam 2)

a) If $\sqrt{n}(T_n - \underline{\mu}) \xrightarrow{D} N_p(\underline{0}, \Sigma)$

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

$A\sqrt{n}(T_n - \underline{\mu}) \xrightarrow{D} N_q(A\underline{0}, A\Sigma A^T)$.

b) Let $\Sigma > 0$ be positive definite.

If $\sqrt{n}(T_n - \underline{\mu}) \xrightarrow{D} N_p(\underline{0}, \Sigma)$ and C_n is a consistent estimator of Σ, then

$D_{T_n}(\underline{\mu}, \frac{C_n}{n}) = n(T_n - \underline{\mu})^T C_n^{-1}(T_n - \underline{\mu}) \xrightarrow{D} \chi^2_p$.

c) If $\Sigma > 0$, $T_n \xrightarrow{p.o.} \underline{\mu}$ and $C_n \xrightarrow{p.o.} \Sigma$, then

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

$= (\underline{x} - \underline{\mu})^T \Sigma^{-1}(\underline{x} - \underline{\mu})$. So

$D^2_{(T_n(1-\alpha)T_n)} \xrightarrow{P} D^2_{\underline{x},1-\alpha}(\underline{\mu}, \Sigma)$ where $P\left(D^2_{\underline{x}}(\underline{\mu}, \Sigma) \le D^2_{\underline{x},1-\alpha}\right) = 1-\alpha$

often $\chi^2_{p,1-\alpha}$

Remark: Let $\underline{Z}_n = \sqrt{n}(\bar{\underline{X}}_n - \underline{\mu}) \Rightarrow N_p(\underline{0}, \Sigma) = \underline{Z}$

Then i) $A\underline{Z} \sim N_k(A\underline{0}, A\Sigma A^T)$

and ii) $A\underline{Z}_n \xrightarrow{D} N_k(\underbrace{A\underline{0}, A\Sigma A^T}_{\text{can't depend on } n}) \xrightarrow{D} A\underline{Z}$.

So i) and ii) behave similarly:

ex) see HW 5 #1.33

ex) Suppose $\underline{X}_i = \begin{pmatrix} X_{i1} \\ X_{i2} \end{pmatrix}$ where

$X_{i1} \sim \text{Pois}(\lambda_1)$ $\perp\!\!\!\perp$ $X_{i2} \sim \text{Pois}(\lambda_2)$,

If $W \sim \text{Pois}(\lambda)$, $E(W) = V(W) = \lambda$.

Find the limiting dist of $\sqrt{n}(\bar{\underline{X}} - \underline{c})$ for appropriate vector $\underline{c}$.

soln) $\sqrt{n}(\bar{\underline{X}} - E(\underline{X})) \xrightarrow{D} N_p(\underline{0}, \Sigma_X)$

$E(\underline{X}) = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \underline{c}$, $\Sigma_X = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ since

$X_{i1} \perp\!\!\!\perp X_{i2} \Rightarrow \text{cov}(X_{i1}, X_{i2}) = 0$. So

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

Unit 1: ... and the bootstrap

1) A large sample $100(1-\delta)\%$ prediction region for $\underline{X}_f$ is a set A_n such that $P(\underline{X}_f \in A_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$. An asymptotically optimal prediction region has volume that converges to the $100(1-\delta)\%$ highest density region.

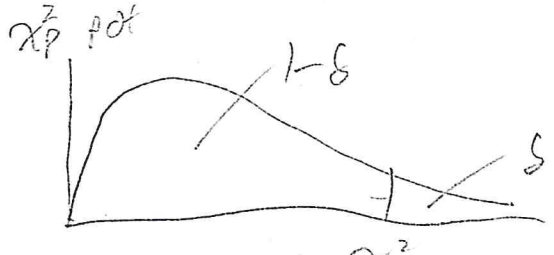
2) Prediction intervals are a special case and we often want $P(\underline{X}_f \in A_n) \to 1-\delta$ as $n \to \infty$.

3) The classical MVN prediction region is

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

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

$$P(X \le \chi^2_{p,1-\delta}) = 1-\delta \text{ if } X \sim \chi^2_p.$$



This region is asymptotically optimal if $\underline{X}_1, \ldots, \underline{X}_n, \underline{X}_f$ are iid $N_p(\mu, \Sigma)$

$\underline{X}_1, \ldots, \underline{X}_n$: training data; $\underline{X}_f$: test data

5) Let $\delta_n = \min(1-\delta, 1-\delta + \frac{10p}{n}), \delta \geq 0.1$

$\min(1-\frac{\delta}{2}, 1-\delta+\frac{10p}{n}) \quad \delta \leq .1.$

If $\delta_n < 1-\delta+0.001$, use $\delta_n = 1-\delta$.

Let $D^2_{(U_n)}$ be the $100\delta_n$ th percentile of $D_i^2(\bar{x}, S)$. The $100(1-\delta)\%$ large sample nonparametric prediction region is

$$\{\underline{z} : D^2_{\underline{z}}(\bar{x}, S) \leq D^2_{(U_n)}\} =$$

$$\{\underline{z} : (\underline{z}-\bar{x})' S^{-1} (\underline{z}-\bar{x}) \leq D^2_{(U_n)}\}.$$

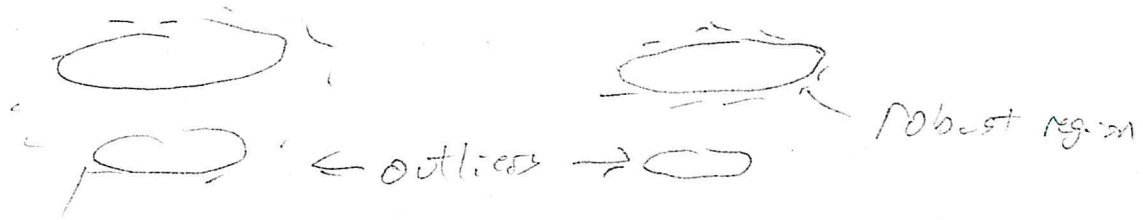
5) This region replaces $\chi^2_{p,1-\delta}$ by $D^2_{(U_n)}$ and is a large sample prediction region if $\underline{x}_1, \ldots, \underline{x}_n$ are iid with nonsingular covariance matrix Σ_x.

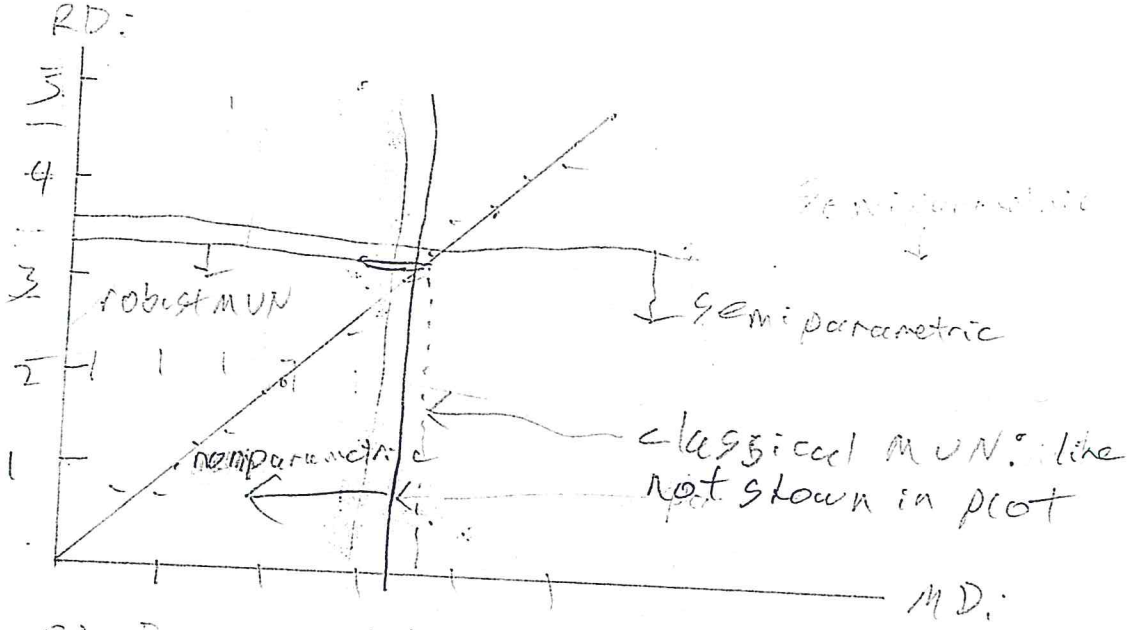
6) The semiparametric prediction region is $\{\underline{z} : D^2_{\underline{z}}(T_{RMVN}, C_{RMVN}) \leq D^2_{(U_n)}(T_{RMVN}, C_{RMVN})\}$

7) Both 4) and 6) are asymptotically optimal on a large class of EC distributions with nonsingular covariance matrices.

8) The semiparametric prediction region can tolerate about $100(1-\alpha_n)\%$ outliers.

$p=2$





ex)

$p=5, n=82$ Nonparametric $MD_i \le D_{(.95\,n)} \approx 3.09$

semiparametric $RD_i \le 3.44$

robust MUN $RD_i \le 3.33$

classical $MD_i \le 3.33$

§4.2 Bootstrap confidence regions

9) Nonparametric bootstrap:

Let data be $\underline{Z}_1, \ldots, \underline{Z}_n$, often iid.

Draw a sample of size n $\underline{Z}_1^*, \ldots, \underline{Z}_n^*$ with replacement from $\underline{Z}_1, \ldots, \underline{Z}_n$

Compute $T_1^* = T_1(X_1^*, \ldots, X_n^*)$, and repeat B times to get a bootstrap sample.

$T_1^*, \ldots, T_B^*$.

10] Ch 2 did this for random variables. See quiz 2 and HW 2.

11] Let $(\bar{T}^*, S^*)$ be the sample mean and covariance matrix computed from $T_1^*, \ldots, T_B^*$.

Let $g_B = \begin{cases} \min\left(1-\frac{\delta}{2}, 1-\delta+10\frac{\delta g}{B}\right), & \delta \leq 0.1 \\ \max\left(1-\delta+.05, 1-\delta+\frac{g}{B}\right) & \delta > 0.1 \end{cases}$

If $1-\delta < 0.999$ and $g_B < 1-\delta+0.001$, set $g_B = 1-\delta$.

12] A large sample $100(1-\delta)\%$ confidence region for θ is a set A_n such that $P(\theta \in A_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$.

13] often want $P(\theta \in A_n) \to 1-\delta$ as $n \to \infty$.

14] If the actual coverage $100(1-\delta_n) < 100(1-\delta)$, the region is liberal. If $100(1-\delta_n) > 100(1-\delta)$, the region is conservative. A region that is 5% conservative is "much better than a region that is 5% liberal.

15] Let $D_{(g_B)}$ be the $100\,g_B$th sample quantile of the D_i.

		Sample 6: