25) take the sample mean of each variable and collect in a vector.

know for E2

ex)

	length X_1	Nasal ht X_2	bigonal X_3	
	168	51	104	1st person
	176	55	106	$n=5$
	179	55	104	
	176	46	114	
	177	48	101	5th person
	$\sum X_{1i} = 876$	$\sum X_{2i} = 255$	$\sum X_{3i} = 529$	

see HW3

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

← sample mean of length

know for E2

ex) Let T be the coordinatewise median, MED(W).

$$T = \begin{pmatrix} 176 \\ 51 \\ 104 \end{pmatrix}.$$

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

26) The R function cov.mcd gives estimator $(\hat{\mu}, \hat{\Sigma})$ that is often useful when outliers are present, but RFCH and RMVN are faster and backed by theory.

27) p107 The jth start $(T_{-1,j}, C_{-1,j})$ is an initial MLD estimator. Then $(T_{0,j}, C_{0,j}) = (\bar{x}_{0j}, S_{0j})$ is the classical estimator computed from the $Cn \approx \frac{n}{2}$ cases with the smallest $D_i(T_{-1j}, C_{-1j})$. Repeat the iteration for k steps resulting in the sequence of estimators $(T_{-1j}, C_{-1j})\ (T_{0j}, C_{0j}), \ldots, (T_{kj}, C_{kj})$. $(T_{kj}, C_{kj}) = (\bar{x}_{kj}, S_{kj})$ the jth attractor. The <u>concentration estimator</u> (T_A, C_A) is the attractor used to obtain the final estimators, $j = 1, \ldots, K$.

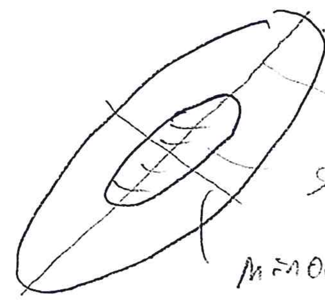
28) p107 $\det(C_{t+1,j}) \leq \det(C_{t,j})$ with equality iff $(T_{tj}, C_{tj}) = (T_{t+1j}, C_{t+1j})$ for $t \geq 0$. So the determinant decreases until convergence.

29) p100 The volume of the hyperellipsoid
$$\{\underset{\sim}{z} : (\underset{\sim}{z} - \bar{\underset{\sim}{x}}_{kj})^T S_{kj}^{-1} (\underset{\sim}{z} - \bar{\underset{\sim}{x}}_{kj}) \leq h^2\} =$$
$$\underbrace{\frac{2\pi^{p/2}}{p\Gamma(\frac{p}{2})}}_{\text{constant}} h^p \sqrt{\det(S_{kj})}.$$
So small volume goes with small determinant.

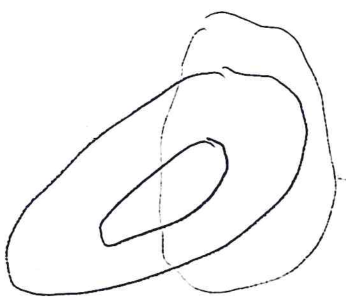
30) For EC distributions, the regions of highest density are $\{\underset{\sim}{z} : (\underset{\sim}{z} - \underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{z} - \underset{\sim}{\mu}) \leq D^2_{h_{\alpha}}(\underset{\sim}{\mu}, \Sigma)\}$ are hyperellipsoids

of Σ. If (T, C) is a consistent estimator of $(\mu, a\Sigma)$ for some $a > 0$, then

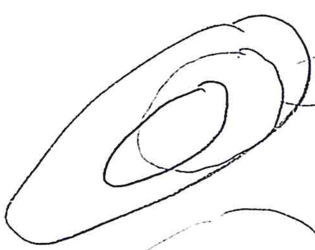
$\{i : D_i^2(T, C) \le MED(D_i^2(T, C))\}$ estimates the highest 50% density region.



31] Using the concentration attractor iterated to convergence seems to estimate a highest density region for EC data + outliers if the attractor region does not contain any outliers.
(If there are 40% outliers, then the highest density region containing $\frac{5}{6}$ of the clean data is estimated.)
$\frac{5}{6}\,60\% = 50\%$



Step 0 Scale ellipsoid to contain half the data take classical estimator



cover half the data
Step 1 new region is a lot better

at convergence, region for attractor seems pretty good

32] Σ determines the shape of the highest density region which is a hyper ellipsoid for EC distributions. (with $g \downarrow$),

iii) An elemental start for MLD uses $p+1$ randomly selected cases J_j. Then $(T_{0j}, C_{0j}) = (\bar{X}_{J_j}, S_{J_j})$ = classical estimator applied to the $p+1$ cases. cov.mcd = FMCD uses 500 randomly chosen elemental starts. For each of the $K = 500$ attractors, find $\det(C_{1,K}), \ldots, \det(C_{500,K})$. Suppose $j = m$ corresponds to the attractor with the minimum determinant. Then (T_{FMCD}, C_{FMCD}) uses the attractor $(T_{m,K}, C_{m,K})$.

B4) p311 The DGK estimator is very simple. Let $(T_{-1}, C_{-1}) = (\bar{X}, S)$ = classical estimator be the start. Then (T_{DGK}, C_{DGK}) is the attractor. Using $K = 10$ concentration steps works well. The DGK estimator has much more outlier resistance than the classical estimator.

B5) p311 The MB (median ball) estimator $(T_{-1}, C_{-1}) = (MED(W), I)$ as the start where $MED(W)$ is the coordinatewise median. Hence (T_0, C_0) is the classical estimator applied to the $c_n \approx \frac{n}{2}$ cases closest to $MED(W)$ in Euclidean distance. Use $K = 5$ concentration steps. This estimator is a highly outlier resistant

can also be ...

ellipsoid for (T_j, C_j)

36) the FCH estimator uses the MB attractor if T_{DGK} has a greater Euclidean distance from $MED(W)$ than half the data (so T_{DGK} is outside of the median ball that contains half of the data). Otherwise FCH uses the DGK or MB estimator that has the smallest determinant $\underbrace{(T_A, C_A)}$.

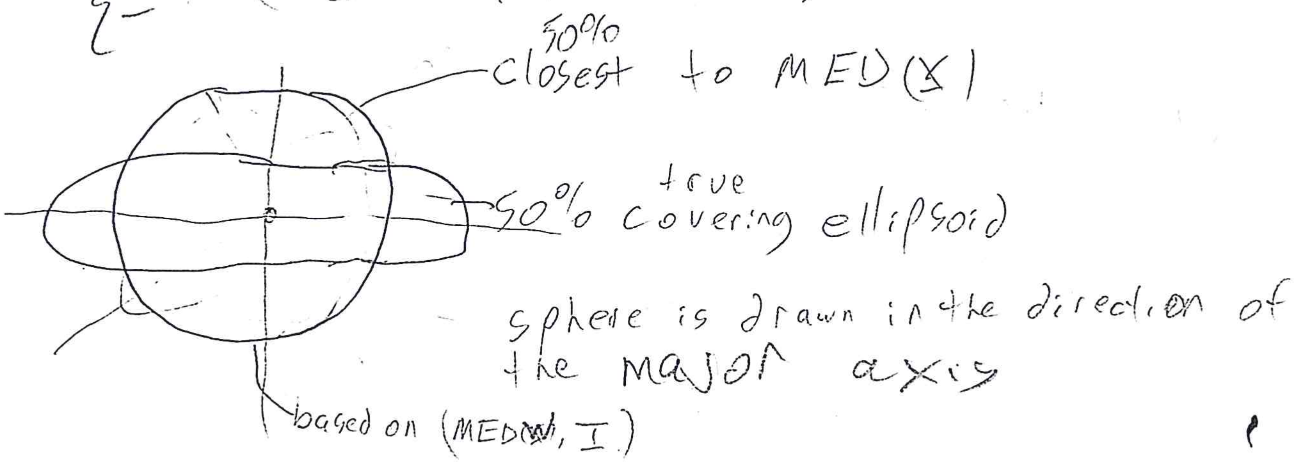
DGK if det DGK = det min
otherwise MB

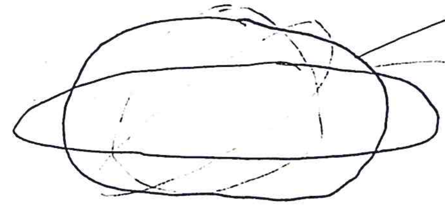
Then $T_{FCH} = T_A$ and

$$C_{FCH} = \frac{MED(D_i^2(T_A, C_A))}{\chi^2_{p,.5}} C_A .$$

37) On a large class of EC distributions, the prob that $(T_A, C_A) = (T_{DGK}, C_{DGK})$ goes to one as $n \to \infty$, and (T_{FCH}, C_{FCH}) is a $\sqrt{n}$ consistent estimator of $(\underline{\mu}, c\Sigma)$ where $c > 0$ and $c = 1$ for MVN data.

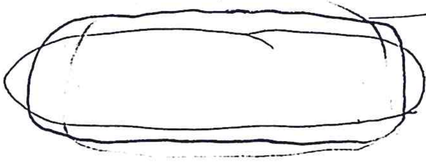
38) The median ball attractor results in ellipsoids $\{\underline{x} \mid (\underline{x} - \underline{T}_{MB})^T C_{MB}^{-1} (\underline{x} - T_{MB}) \le d^2\}$ that are "too short" in the major axis and "too fat" in the minor axis compared to $\{\underline{x} \mid (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \le d^2\}$.





(bMP, SOS)

too fat on the minor axis
too short on the major axis



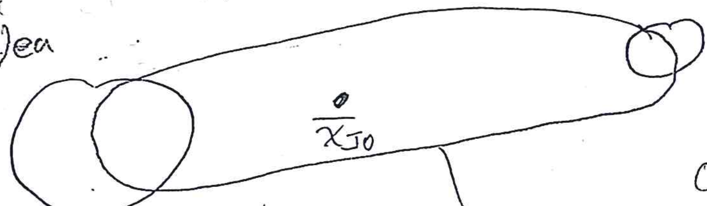
based on
(T_{KSD}, C_{K}) = (T_{MB}, C_{MB})
better but still biased if
(k concentration steps are used,)
(It is not known whether MB is biased or not) if concentration is iterated to convergence)

21) For ...
... concentration
the ... estimator is ...
By 19... MB ...

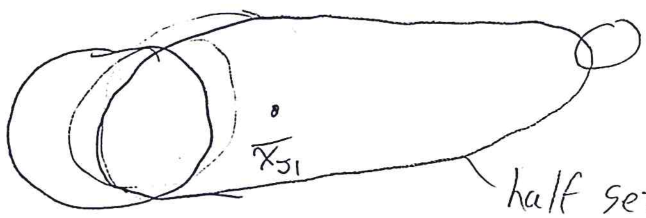
39) When outliers are present that the DGK estimator can't detect, usually $\det(C_{MB}) < \det(C_{DGK})$ so the FCH estimator is highly outlier resistant.

40) Even if the 50% of cases with the smallest distances based on the start contains outliers, the attractor may use a half set without outliers.

44) Idea



If the number of outliers in the half set $J < 25\% n$, the $\bar{x}_J$ is closer to the bulk of the data than to the outliers. After a concentration step, fewer outliers and more clean cases will be used (if the outliers are far enough away). After k concentration steps, the half set may be clean.



clean data "won the tug of war"
HW 7C illustrates this phenomenon.

half set J_k contains no outliers.

42) The ~~MCD~~ FCH estimator can give good results even when nearly 50% of the cases are outliers.

43) A scatterplot matrix of the distances from cov.mcd, FCH, DGK and the median ball estimator can be useful. HW 7B makes DD plots based on these 4 estimators.

---

44) p. 116 RFCH and RMVN are reweighted versions of FCH. They are $\sqrt{n}$ consistant estimators of $(\mu, c_i \Sigma)$ where $c_i = 1$ for MVN data, $c_1 = c_{RFCH}$, $c_2 = c_{RMVN}$.