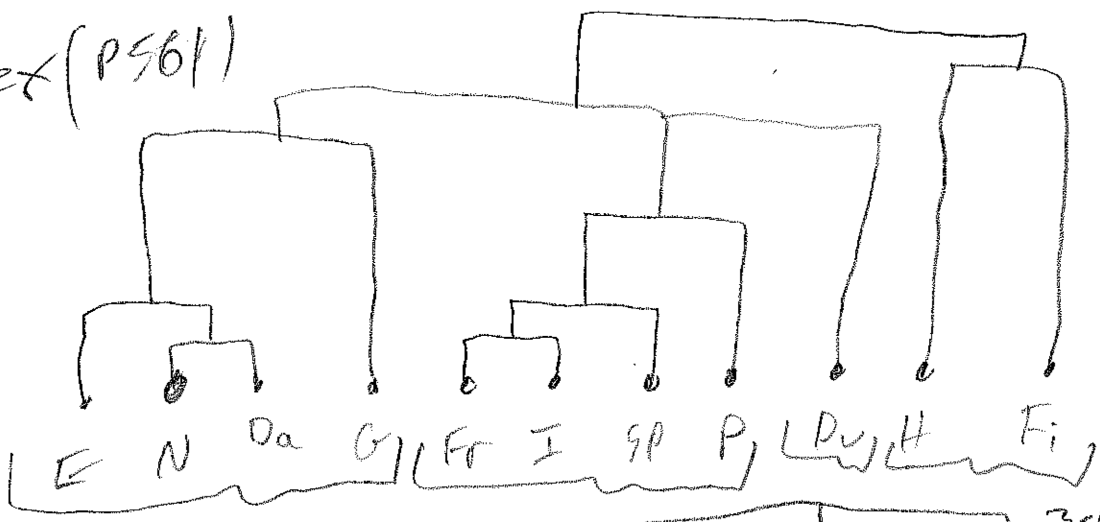
MV55

dendrogram to select the reasonable clusters of cases.

ex(p561)



Complete linkage

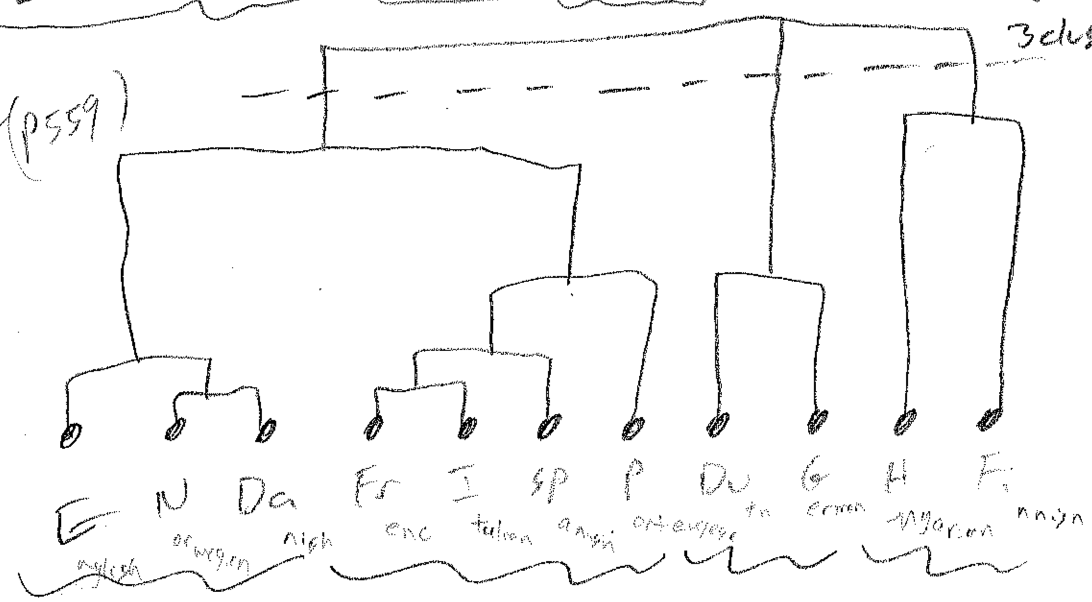
* vertical height represents closeness of clusters

3 clusters

could cluster with a horizontal line

Single linkage

ex(p559)



11 languages

could use 5 clusters of languages

could use 3 clusters

This example considers the numbers 1-10 in ten languages.

one <u>en</u> <u>en</u> <u>een</u> <u>ein</u> un uno uno jeden <u>egy</u> yksi | 95.5

Look at the first letter of each word. Two different languages are concordant if they have the same 1st letter.

Concordant 1st letters

	E	N	Da	Du	G	F	Sp	I	P	H	Fi
E	10										
N	8	10									
Da	8	9	10								
Du											
G											
F											
Sp											
I											
P											
H											
Fi	1	1	1					1		2	10

10 − concordant matrix
= distance matrix
so 10 → 0
9 → 2
1 → 9 etc

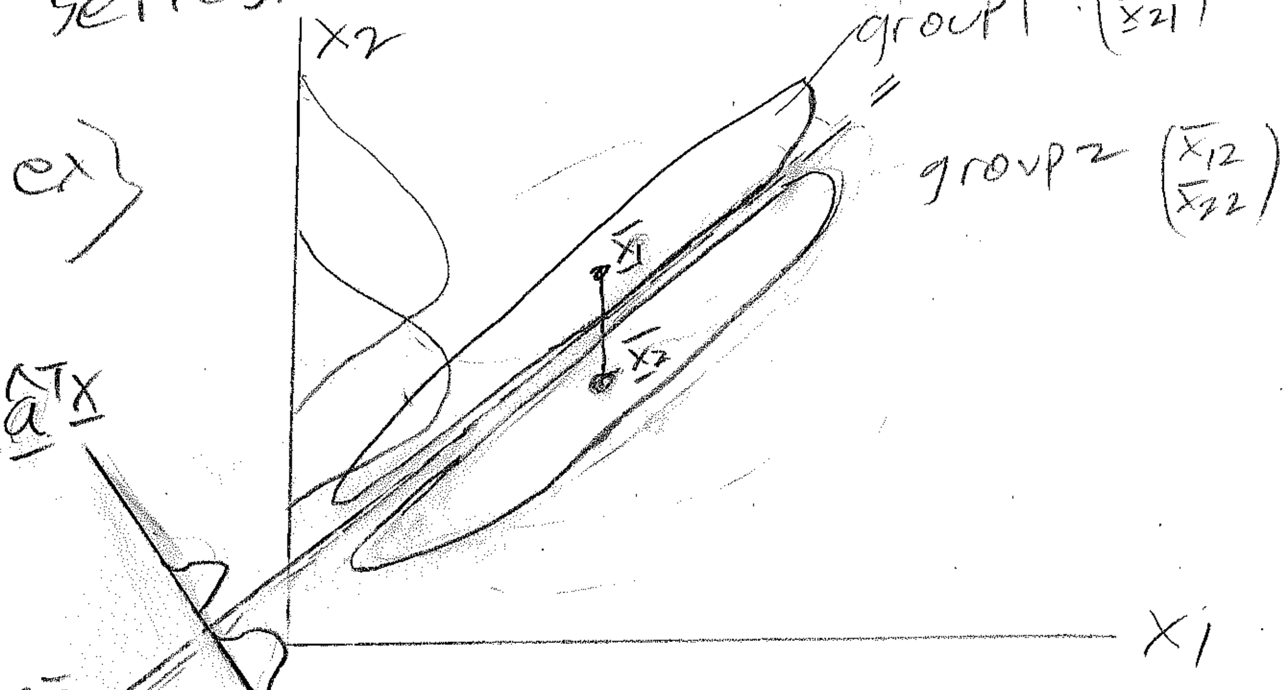
7) Dendrograms are hard to use with large data sets. Need an automated method to choose the final k clusters for large data sets

8) Often take training data with known groups to check the cluster analysis.

9) For cluster analysis, we try to find groups, but the groups are unknown for both training data and test data. Sometimes, there is no test data.

A Fairly recent techniques for multivariate analysis and regression are called envelope methods. Dr Samadi uses the technique for multivariate time series.

ex)



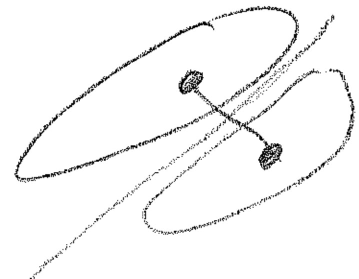
$\bar{x}_{11} = \bar{x}_{12}$, $\bar{x}_{21} \neq \bar{x}_{22}$

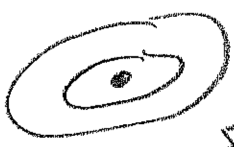
Need n_1, n_2 large for "usual" testing $\underline{\mu}_1 = \underline{\mu}_2$

If you project the data on the vector $\hat{a}$, using $\hat{a}^T x$, there is good discrimination between the two groups. Projection on $\hat{b}$ has nearly no discrimination power. So $\hat{a}$ is the "material part" while $\hat{b}$ is the "immaterial part".

For this ex, $\hat{a} \approx \hat{e}_2$ from PCA works better than $\hat{a} \approx \bar{x}_1 - \bar{x}_2$.

Often $\hat{a} \approx \bar{x}_1 - \bar{x}_2$ is good at discrimination.





$\mu_1 = \mu_2$

$\bar{x}_1 - \bar{x}_2 \approx 0$ is bad

There are envelope methods for multivariate linear regression, both for the responses $Y_1, \ldots, Y_m$ and the predictors $x_1, \ldots, x_p$.

Inference for covariances

see 5.3.4

1) a) A sequence of random variable W_n is bounded in probability, written $W_n = O_P(1)$ (big-O-P-1), if for every $\varepsilon > 0$ $\exists$ constants D_ε and N_ε $\ni$ $P(|W_n| \le D_\varepsilon) \ge 1-\varepsilon$ $\forall\, n \ge N_\varepsilon$. Also $W_n = O_P(X_n)$ if

$$\left|\frac{W_n}{X_n}\right| = O_P(1).$$

little-O-P-1

b) The sequence $W_n = o_P(1)$ if $W_n \xrightarrow{P} 0$. The sequence $W_n = o_P(n^{-\delta})$ if $n^\delta W_n = o_P(1)$,

2) Let $A_n = (a_{ij}(n))$ be an $r \times c$ random matrix. $A_n = O_P(X_n)$ if $a_{ij}(n) = O_P(X_n)$ for $1 \le i \le r$ and $1 \le j \le c$.

$A_n = o_p(X_n)$ if $a_{ij}(n) = o_p(X_n)$ for $1 \le i \le r$ and $1 \le j \le c$.

3) If $\underline{Z}_n = O_p(n^{-\delta})$ and $0 < \varepsilon < \delta$,

then $\underline{Z}_n = o_p(n^{-\varepsilon})$ so $n^{\varepsilon} \underline{Z}_n \xrightarrow{P} 0$.

4) If $\underline{Z}_n \xrightarrow{D} \underline{Z}$, then $\underline{Z}_n = O_p(1)$,

If $n^{\delta}(\underline{Z}_n - \underline{\theta}) \xrightarrow{D} \underline{Z}$, then $\underline{Z}_n = O_p(n^{-\delta})$.

5) a) If $W_n = O_p(1)$ and $X_n = O_p(1)$, then $W_n + X_n = O_p(1)$ and $W_n X_n = O_p(1)$,

so $O_p(1) + O_p(1) = O_p(1)$ and $O_p(1) O_p(1) = O_p(1)$.

b) If $W_n = O_p(1)$ and $X_n = o_p(1)$, then

$W_n + X_n = O_p(1)$ and $W_n X_n = o_p(1)$,

so $O_p(1) + o_p(1) = O_p(1)$ and $O_p(1)\, o_p(1) = o_p(1)$.

c) If $X_{in} = O_p(n^{-\delta})$, $i = 1, \ldots, n$

then $\sum_{i=1}^{n} X_{in} = n\, O_p(n^{-\delta})$ and $\frac{\sum_{i=1}^{n} X_{in}}{n} = O_p(n^{-\delta})$, Suppose $x_{in} = O_p(n^{-\delta_i})$ $i = 1, \ldots, n$

If $Y_{in} \equiv X_{jn}$ where $\delta_j = \min(\delta_1, \ldots, \delta_n)$, $\delta_i > 0$

then $\sum_{i=1}^{n} Y_{in} = n X_{jn} = n \, O_p(n^{-\delta_j})$

and $\frac{1}{n} \sum_{i=1}^{n} Y_{in} = X_{jn} = O_p(n^{-\delta_j})$.

cancellation could occur so that better bounds can be found

6} Let $\underset{p \times p}{A}$ be symmetric. vech(A) stacks the elements of A on and below the diagonal from left column to right column.

ex} $A = \begin{bmatrix} 1 & & \\ .2 & 4 & \\ .3 & .5 & 6 \end{bmatrix}$

$\text{vech}(A) = (1, .2, .3, 4, .5, 6)^T$.

7} See E3 rev 138)

Let $\underset{\sim}{c} = [\sigma_{ij}]$ be a vector of distinct elements of Σ_x. Let the estimator $\hat{\underset{\sim}{c}} = [\hat{\sigma}_{ij}]$

have the same ordering as $\underset{\sim}{c}$.

The largest subset of distinct elements corresponds to vech(Σ_x) =

$$(\sigma_{11}, \ldots, \sigma_{1p}, \sigma_{22}, \ldots, \sigma_{2p}, \ldots, \sigma_{p-1,p-1}, \sigma_{p-1,p}, \sigma_{pp})^T.$$

Could use $\underset{\sim}{c} = (\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp})^T$

(the variances), $c = \sigma_{11} = \sigma_1^2$, etc.

For random variables $x_1, \ldots, x_p$, use notation such as $\overline{x}_j$ = sample mean of the x_j, $\mu_j = E(x_j)$, and $\sigma_{jk} = \text{Cov}(x_j, x_k)$.

Let $\hat{\underset{\sim}{c}} = [\tilde{\sigma}_{jk}]$, $n\hat{c} = \sum_{i=1}^{n} [(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)]$

$$= \sum_{i=1}^{n} [n \tilde{\sigma}_{jk}] = \sum_{i=1}^{n} \underset{\sim}{z}_i \quad \text{where}$$

$\underset{\sim}{z}_i = [(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)]$ and

$\underline{w}_i = [(x_{ij} - \mu_j)(x_{ik} - \overline{x}_k)]$.

Let $\text{Cov}(\underline{w}_i) = \Sigma_{\underline{d}}$ to avoid confusion with a $\Sigma_{\underline{w}}$ used later.

Then $E(\underset{\sim}{w}_i) = E(\overline{\underset{\sim}{w}}_n) = \underset{\sim}{c}$.

8) Assume the cases $\underset{\sim}{x}_i$ are iid and that $Cov(\underset{\sim}{w}_i) = \Sigma_{\underset{\sim}{d}}$. Using the notation in 7) with $\underset{\sim}{c}$ a $k \times 1$ vector,

i) $\sqrt{n}(\tilde{\underset{\sim}{c}} - \underset{\sim}{c}) \xrightarrow{D} N_k(\underset{\sim}{0}, \Sigma_{\underset{\sim}{d}})$

ii) $\sqrt{n}(\hat{\underset{\sim}{c}} - \underset{\sim}{c}) \xrightarrow{D} N_k(\underset{\sim}{0}, \Sigma_{\underset{\sim}{d}})$

iii) $\hat{\Sigma}_{\underset{\sim}{d}} = \hat{\Sigma}_{\underset{\sim}{z}} + O_P(n^{-1/2})$ and $\tilde{\Sigma}_{\underset{\sim}{d}} = \tilde{\Sigma}_{\underset{\sim}{z}} + O_P(n^{-1/2})$,

So $\hat{\Sigma}_{\underset{\sim}{z}}$ and $\tilde{\Sigma}_{\underset{\sim}{z}}$ are $\sqrt{n}$ consistent estimators of $\Sigma_{\underset{\sim}{d}}$.

Proof} i) Since the $\underset{\sim}{x}_i$ are iid, the $\underset{\sim}{w}_i$ are iid with $E(\underset{\sim}{w}_i) = \underset{\sim}{c}$ and $Cov(\underset{\sim}{w}_i) = \Sigma_{\underset{\sim}{d}}$. Hence by the MCLT, $\sqrt{n}(\overline{\underset{\sim}{w}}_n - \underset{\sim}{c}) \xrightarrow{D} N_k(\underset{\sim}{0}, \Sigma_{\underset{\sim}{d}})$.

Then $n\tilde{\underset{\sim}{c}} = \sum_i [(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)] \overset{\leftarrow \text{vector}}{=}$

$$\sum_i [(x_{ij} - \mu_j + \mu_j - \overline{x}_j)(x_{ik} - \mu_k + \mu_k - \overline{x}_k)] =$$

$$\underbrace{\sum_i [(x_{ij} - \mu_j)(x_{ik} - \mu_k)]}_{\underset{\sim}{w}_i} + \underbrace{\sum_i [(x_{ij} - \mu_j)(\mu_k - \overline{x}_k)]}_{[n(\overline{x}_j - \mu_j)(\mu_k - \overline{x}_k)] = -\underset{\sim}{a}_n} +$$

$$\underbrace{\sum_i \left[(\mu_j - \bar{x}_j)(x_{ik} - \mu_k)\right]}_{-\underline{a}_n = \left[n\,(\mu_j - \bar{x}_j)(\bar{x}_k - \mu_k)\right]} + \underbrace{\sum_i \left[(\mu_j - \bar{x}_j)(\mu_k - \bar{x}_k)\right]}_{\left[n\,(\mu_j - \bar{x}_j)(\mu_k - \bar{x}_k)\right] = \underline{a}_n}$$

$= \left(\sum_i \underline{w}_i\right) - \underset{\sim}{a}_n$ where

$$\underline{a}_n = \left[\underbrace{\sqrt{n}\,(\bar{x}_j - \mu_j)}_{O_p(1) \text{ by CLT}}\ \underbrace{\sqrt{n}\,(\bar{x}_k - \mu_k)}\right] = O_p(1).$$

So $n\,\underset{\sim}{\tilde{c}} = \sum_i \underset{\sim}{w}_i - \underline{a}_n$ and

So $\sqrt{n}\,\underset{\sim}{\tilde{c}} = \dfrac{\sqrt{n}\sum_i w_i}{\sqrt{n}\sqrt{n}} - \dfrac{a_n}{\sqrt{n}}$ and

$$\sqrt{n}\left(\underset{\sim}{\hat{c}} - \underset{\sim}{c}\right) = \underbrace{\sqrt{n}\left(\overline{w}_n - \underset{\sim}{c}\right)}_{\xrightarrow{D} N_k(0, \Sigma_d)} - \underbrace{\frac{a_n}{\sqrt{n}}}_{\to 0} \xrightarrow{D} N_k(\underline{0}, \Sigma_{\underset{\sim}{d}})$$

iii) $\underline{w}_i = \left[(x_{ij} - \mu_j)(x_{ik} - \mu_k)\right] =$

$\left[(x_{ij} - \bar{x}_j + \bar{x}_j - \mu_j)(x_{ik} - \bar{x}_k + \bar{x}_k - \mu_k)\right] =$

$$\underbrace{\left[(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)\right]}_{\underset{\sim}{z}_i} +$$

$$\underbrace{\left[(x_{ij}-\bar{x}_j)(\bar{x}_k-\mu_k)\right] + \left[(\bar{x}_j-\mu_j)(x_{ik}-\bar{x}_k)\right]}_{\underset{\sim}{a}_i} + \left[(\bar{x}_j-\mu_j)(\bar{x}_k-\mu_k)\right]$$

Σ same

$$\text{so } \underline{\bar{w}} = \underline{\bar{z}} + \underline{0} + \underline{0} + \left[(\bar{x}_j-\mu_j)(\bar{x}_k-\mu_k)\right]$$

$$\text{so } \underline{w}_i - \underline{\bar{w}} = \underline{z}_i - \underline{\bar{z}} + \underset{\sim}{a}_i \quad \text{where}$$

$$\underline{a}_i = \left[(x_{ij}-\bar{x}_j)(\bar{x}_k-\mu_k)\right] + \left[(\bar{x}_j-\mu_j)(x_{ik}-\bar{x}_k)\right] \quad \text{and}$$

$$\sqrt{n}\,\underline{a}_i = \left[\underbrace{(x_{ij}-\bar{x}_j)}_{O_p(1)}\ \underbrace{\sqrt{n}(\bar{x}_k-\mu_k)}_{O_p(1)}\right] + \left[\underbrace{\sqrt{n}(\bar{x}_j-\mu_j)}_{O_p(1)}\ \underbrace{(x_{ik}-\bar{x}_k)}_{O_p(1)}\right] = O_p(1).$$

$$\therefore\ \underset{\sim}{a}_i = O_p(n^{-1/2}). \quad \text{Thus}$$

$$\tilde{\Sigma}_{\underset{\sim}{d}} = \frac{1}{n}\sum (\underline{w}_i - \underline{\bar{w}})(\underline{w}_i - \underline{\bar{w}})^T =$$

$$\frac{1}{n}\sum_i (\underline{z}_i - \underline{\bar{z}} + \underline{a}_i)(\underline{z}_i - \underline{\bar{z}} + \underset{\sim}{a}_i)^T =$$

$$\frac{1}{n}\sum_i (\underline{z}_i-\underline{\bar{z}})(\underline{z}_i-\underline{\bar{z}})^T + \frac{1}{n}\sum_i \underbrace{(\underline{z}_i-\underline{\bar{z}})\,\underline{a}_i^T}_{O_p(n^{-1/2})} + \frac{1}{n}\sum_i \underbrace{\underline{a}_i(\underline{z}_i-\underline{\bar{z}})^T}_{O_p(n^{-1/2})}$$

$$+ \frac{1}{n}\sum_i \underbrace{\underline{a}_i\underline{a}_i^T}_{O_p(n^{-1})} = \underbrace{\frac{1}{n}\sum_i (\underline{z}_i-\underline{\bar{z}})(\underline{z}_i-\underline{\bar{z}})^T}_{\tilde{\Sigma}_z} + O_p(n^{-1/2})$$

by 5] c).

9] So $\sqrt{n}\,(\hat{\underset{\sim}{c}} - \underset{\sim}{c}) \xrightarrow{D} N_k(\underset{\sim}{0}, \Sigma_d)$

and $\hat{\Sigma}_d$ is a $\sqrt{n}$ consistent estimator

of Σ_d. Need $n \geq 10k$ and

often much larger, but do not

need $n > p$, since $\underset{k \times 1}{\hat{\underset{\sim}{c}}}$, $\underset{k \times 1}{\underset{\sim}{c}}$ and $\underset{k \times 1}{\underset{\sim}{z}_i}$

only depend on the k distinct entries

in Σ.

10] For $k = 1$, can get CIs for

σ_{ij}, eg for $\sigma_1^2, \ldots, \sigma_p^2$.

These results lead into high dimensional

multivariate statistics.

11] A data set is low dimensional

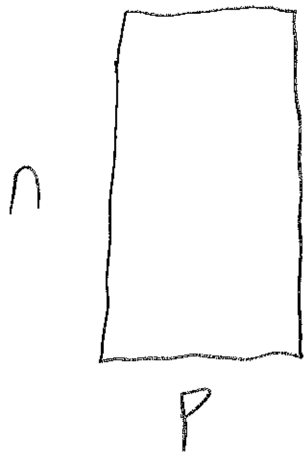
if $n \geq Jp$ with $J \geq 5$.

A data set is high dimensional if

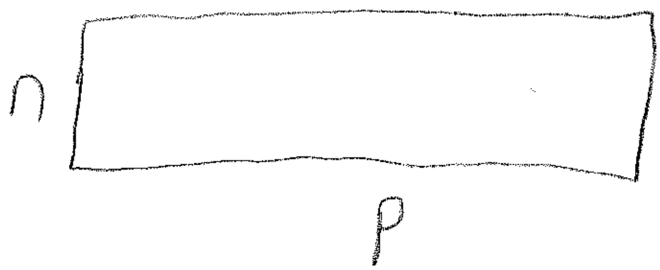
$n \leq 5p$. Note that $n << p$ is possible.

eg $n = 100$, $p = 1$ million.

The multivariate data set $\underset{p \times 1}{\boldsymbol{x}_1}, \ldots, \boldsymbol{x}_n$ has p measurements.



classical statistics



high dimensional statistics

2) High dimensional statistics is perhaps the hottest research topic since 2000.

3) There are 2 important techniques for HD statistics.

a) model selection or variable selection:
select k of the p predictors with
$n \geq 5k$ (often much bigger)

In HW 9B (for the pottery data)
$n = 28$, $p = 20$ but crude

variable selection got 4 predictors $x2, x3, x5$ and $x8$ for perfect classification of training data.

b) Suppose $\sqrt{n}(\underset{p\times 1}{\hat{\underline{\theta}}-\underline{\theta}}) \xrightarrow{D} N_p(\underline{0}, \Sigma)$

where $\hat{\Sigma} \xrightarrow{P} \Sigma$ and $\hat{\Sigma}$ can be computed

($\hat{\Sigma}$ often $= \hat{C}^{-1}$ can not be computed if there is matrix inversion).

Then $\hat{\underline{\theta}} \sim N_p(\underline{\theta}, \frac{\hat{\Sigma}}{n})$ and can get tests and CIs for k components of θ, eg CIs for $\theta_1, \ldots, \theta_p$.

This inference occured for the covariance matrix $\sqrt{n}(\tilde{\underset{\sim}{c}} - \underset{\sim}{c}) \xrightarrow{D} N_k(\underset{\sim}{0}, \Sigma_{\underset{\sim}{d}})$.

4) Often $\sqrt{n}(\hat{\underline{\theta}}-\underline{\theta}) \xrightarrow{D} N_p(\underline{0}, \Sigma)$. Then

$\hat{\theta}_i - \theta_i = O_p(n^{-1/2}) \propto \frac{1}{\sqrt{n}}$

So $\|\hat{\underline{\theta}} - \underset{\sim}{\underline{\theta}}\|^2 = \sum_{i=1}^{p}(\hat{\theta}_i - \theta)^2 \propto \frac{p}{n}$