Big Data Statistics. STAT3017 / STAT7017.

Dr Dale Roberts.

* **Course Outline** (Available on Wattle)
    - About me.
    - Assessments.
    - Course schedule.

* **Structure.**
    - 2 hours lectures.
    - 1 hour workshop / computer lab. (start next week).

* **Material.**
    - Lecture notes (Handwritten).

      ↳ scanned & placed on Wattle.

    - PDFs of research papers. → Wattle.
    - Extracts from books.
    - R codes.

What is BIG DATA?

Wikipedia: "data sets that are so large or complex that traditional data processing applications are inadequate".

Gartner 2012: 3Vs

High Volume: "data not sampled"

Velocity: "real-time"

Variety: "draws from text, images, ... video".

I personally **HATE** these definitions, because:

- Data processing / computing is focus.
  → What happens in 10 years when this isn't a problem anymore? (Moore's law)
- Doesn't properly capture the true (and timeless) difference to "small data".

Q: Are large sample sizes really the problem?

"Volume"

1000 kilobyte

1000^2 megabyte

1000^3 gigabyte

1000^4 terabyte

1000^5 petabyte

1000^6 exabyte

$\vdots$

Big data?

Large sample theory is basis for classic statistics.

$X_i \sim F$ iid. for $i=1,\dots,n$ $\quad \mathbb{E}X_i = \mu$.

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$$

Law of large numbers $\overline{X}_n \to \mathbb{E}X$ as $n \to \infty$

Central limit theorem $\sqrt{n}\left(\overline{X}_n - \mathbb{E}X\right) \to N(0, \cdot)$

Big data should only reaffirm very classic theory!

Q: Is real-time data a problem?

Yes, but most data sets are not "real-time".

There is interesting theory here for <u>streaming data</u>

ONLINE LEARNING, etc.

(I will not cover this topic this semester.)

Q: Is data variety a problem?

Not really. The topic of <u>multivariate analysis</u>

has existed since early 1900s.

Multivariate analysis. Given a sample $\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_n$ of random obs of dimension p.

$$\mathbf{x}_i = [\, x_i^{(1)} \; x_i^{(2)} \; \dots \; x_i^{(p)} \,]$$ (or transposed version)

Methods such as PCA have been available since early 1900s.

$$\text{Obs Gaussian:} \quad \left.\begin{array}{l}\text{Student's T-test} \\ \text{Fisher's test.} \\ \text{ANOVA.}\end{array}\right\} \text{Non-asymptotic methods.}$$

Non-Gaussian: (non-asympt.) results are hard to obtain

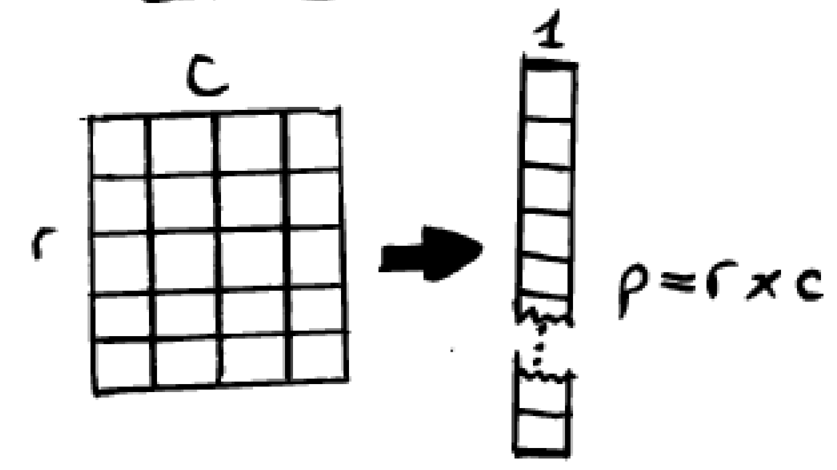
→ limiting theorems based on model statistics.

Typically desired under assumptions:
p fixed $\quad n \to \infty$ "large sample theory".

Classic MVA $p < 10$.

New challenge: BIG DATA!

	p	n	p/n
Portfolio	~ 50	500	0.1.
Climate survey	320	600	0.21.
Speech analysis	$a \times 10^2$	$b \times 10^2$	~ 1.
Face database	1440	320	4.5.
Micro-array	1000	100	10.



I shall define BIG DATA as "data whereby the classic statistical paradigm <u>no longer applies.</u>"

<u>classic paradigm:</u>

- dimension p is small compared to the sample size n.

- asymptotic theory assumes n increases while dimension p remains fixed.

- At time t, we have all the data necessary for our analysis, ie. the <u>batch</u> case.

<u>No</u> longer applies means:

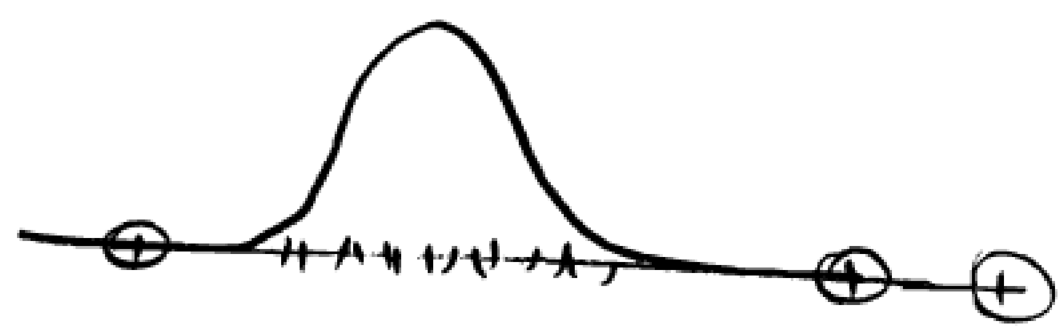
- gives incorrect results.
- bad approximation.
- incorrect hypothesis rejection.
- etc.

Unique features of big data:

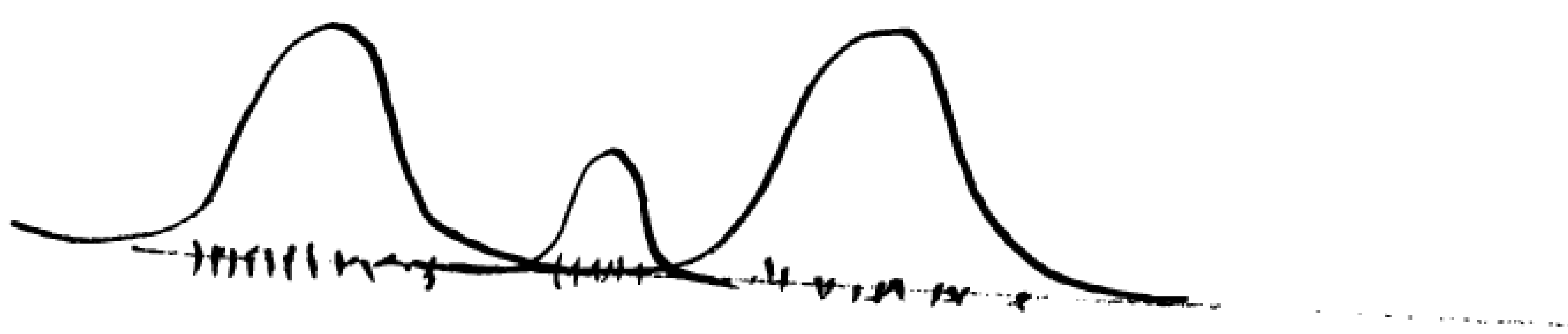
[Fan, Han, Liu; 2014]
and references therein

(Quick overview as I haven't presented notation yet).

Heterogeneity: With small data, data points from subpopulations are considered 'outliers'.



With large data sets, subpopulations might be large.



⟹ Mixture of Gaussians?

Noise accumulation: Errors accumulate when a decision or or prediction rule depends on a large number of parameters.

This effect becomes worse as the dimension increases. and may dominate the true signal.

(See Fig 1)

Spurious correlation:

High dimensionality can cause spurious correlations. That is, many uncorrelated random variables may have high sample correlation.

(See Fig 2)

Incidental endogeneity

In regression setting,

$$Y = \sum_{i=1}^{p} X_i + \varepsilon$$

'engogeneity' means some features (predictors) correlate with the residual noise ε.

That the residual noise ε is uncorrelated with all features is crucial. "Exogenous assumption"

$$\mathbb{E}[\varepsilon X_i] = 0 \quad \text{for } i = 1, \cdots, p$$

Easily violated in high-dimensions.

Aim of the course.

Go from classic $\longrightarrow$ cutting-edge

- High dimensional ($p \approx n$ large or $p >> n$)
- ~~Streaming (sequentially revealed)~~

Not this semester.

We need to understand the classic case to see why new approaches are better.

This is an active area of research: lots of open questions and new applications to find.

Fundamental idea: Study Random Matrices.

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{12} & X_{22} & & \\ \vdots & & \ddots & \\ X_{p1} & & & X_{pp} \end{bmatrix} \qquad X_{ij} : \Omega \to \mathbb{R} \ (\text{or } \mathbb{C}).$$

What is a Random Matrix?

[Diaconis '05].

"Everyone knows" that a random variable is just a measurable function from our sample space Ω.

$$X : \Omega \to \mathcal{S} \qquad \mathcal{S} = \mathbb{R},\ \mathbb{R}^n, \ldots$$

Take $\mathcal{S} = \mathbb{R}^{n \times n}$ i.e. $n \times n$ matrices with real entries.

"That's not what it means to people working in probability"

Think about <u>picking</u> a matrix (with certain properties) at random with a certain probability.

Eg. Pick a random covariance matrix.

$\boxed{\text{Matrix Properties}} + \boxed{\text{Randomness}} =$ Interesting Maths!

RMT Quantum mechanics 40's – 50's.

- Energy levels of a system are described by eigenvalues of a Hermitian operator $\mathbb{A}$ on a Hilbert space.
- Computationally you can't work on ∞-dim objects.
  - $\Rightarrow$ discretisation & truncation: keep only parts that are important to the problem under consideration
  - $\Rightarrow$ $\mathbb{A}$ finite but large random linear operator.

- Semicircular law for Gaussian (or Wigner) matrix [Wigner 1955; 1958]
  - $\Rightarrow$ [Arnold 1967; 1971] [Grenander 1963].

- Gaussian Wishart matrices (sample covariance matrices).
  [Marčenko/Pastur 1967] [Pastur 1972; 1973].
  $\Rightarrow$ Marčenko-Pastur law.

- Asymptotic theory of large sample covar matrices.
  [Bai et al 1986] [Grenander & Silverstein 1977]
  [Johnsson 1982] ...
  Multivariate Fisher matrices (QR^{-1}) $Q, R \perp\!\!\!\perp$ sample covar matrices.

- Recently 2nd-order theory: CLT for linear spectral statistics, limit dist spectral spacings, extreme eigenvalues.

Sample covariance matrices.

$X_1, X_2, \cdots X_n$ sample of random obs. dimension p.

Population covariance matrix; $\Sigma = \mathrm{cov}(X_i)$

Sample covariance matrix: $S_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^*$

Sample mean: $\bar{X} = \frac{1}{n} \sum X_i$

Most results in MVA rely on S_n: PCA, Canonical correlation analysis, multivariate regression, one-sample or two-sample hypothesis testing, factor analysis.

$\Rightarrow$ Understanding asymptotic properties of S_n extremely important in data analysis when p becomes large wrt. sample size n.

- Generalised Variance & multiple correlation coefficient.

  $\Rightarrow$ overall measure of dispersion of the data. σ_i^2 measures X_i

  all variables together: generalised variance. "measure of scatter".

p becomes large $\Rightarrow$ "BIG DATA."

RMT will become our tool to understand what is happening.

Review of some Matrix Algebra.

A complex number is a number of the form $a+ib$ where i satisfies $i^2 = -1$.

$$Re[a+ib] = a \qquad Im[a+ib] = b.$$

The complex conjugate of $z = a+ib \in \mathbb{C}$ is $\bar{z} := a - ib$

if A is a $m \times n$ matrix with complex entries, then the $n \times m$ matrix A^* is called the conjugate transpose and is defined as $[A^*]_{ij} := \overline{A_{ji}}$ or $A^* := (\bar{A})' = \overline{(A')}$

The matrix $A = (a_{ij})$ is Hermitian if it is square with $a_{ij} \in \mathbb{C}$ such that $A = A^*$. The matrix A is symmetric if $A = A'$ and orthogonal if $A'A = AA' = I$ where I is the identity matrix, equivalently $A' = A^{-1}$. A complex square matrix is called unitary if $A^* A = A A^* = I$.

The product AB of $m \times n$ matrix $A = (a_{ij})$ and $n \times k$ matrix $B = (b_{ij})$ is the $m \times k$ $C = (c_{ij})$ where

$$c_{ij} = \sum_{l=1}^{n} a_{il} b_{lj} \quad \text{for } i = 1, 2, \ldots, m \text{ \& } j = 1, 2, \ldots, k.$$

The transpose of a matrix A is A' such that $[A']_{ij} = [A]_{ji}$

The trace of a $k \times k$ matrix $A = (a_{ij})$ is $Tr(A) = \sum_{l=1}^{k} a_{ll}$

The <u>determinant</u> of A, denoted $|A|$ or $\det(A)$, is the scalar $|A| = a_{11}$ if $k=1$ or $|A| = \sum_{j=1}^{k} a_{1j} |A_{1j}| (-1)^{1+j}$ if $k>1$ where A_{1j} is the $(k-1)\times(k-1)$ matrix obtained by deleting the first row and j'th column of A.

For $k \times k$ matrices A and B, constant $c \in \mathbb{R}$, we have:

$$(A+B)' = A' + B'$$

$$(AB)' = B'A'$$

$$\det(A') = \det(A)$$

$$(A')^{-1} = (A^{-1})'$$

$$tr(cA) = c\, tr(A)$$

$$tr(A \pm B) = tr(A) \pm tr(B)$$

$$tr(AB) = tr(BA)$$

$$tr(B^{-1}AB) = tr(B)$$

$$(AB)^{-1} = B^{-1}A^{-1}$$

$$tr(AA') = \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij}^2$$

$$\det(AB) = \det(A)\det(B)$$

$$\det(cA) = c^k \det(A)$$