STAT7017 Big Data Statistics	**Semester 2 2018**

Lecture 1: 23 July

Lecturer: Dr Dale Roberts — *Scribes: Rui Qiu*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

- Course Outline (available on Wattle)
  - About me.
  - Assessments. (15+15+15+55)
  - Course schedule.
- Structure
  - 2 hours lectures.
  - 1 hour workshop / computer lab. (start next week).
- Material
  - Lecture notes (Handwritten), scanned and placed on Wattle.
  - PDFs of research papers.
  - Extracts from books.
  - R codes.

Q: 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"
- High Velocity: "real-time"
- High 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?

In terms of "volume"

1000	kilobyte
1000^2	megabyte
1000^3	gigabyte
1000^4	terabyte
1000^5	petabyte
1000^6	exabyte
...	...

Starting from *gigabyte*, big data?

Large sample theory is basis for classic statistics:

$X_i \sim$ F iid, for $i = 1, \ldots,$

$$\begin{aligned} \mathbb{E}X_i &= \mu \\ \bar{X}_n &:= \frac{1}{n}\sum_{i=1}^{n} X_i \end{aligned} \tag{1.1}$$

Law of large numbers $\bar{X}_n \longrightarrow \mathbb{E}X$ as $n \longrightarrow \infty$. (sample mean converges to population mean)

Central limit theorem $\sqrt{n}(\bar{X}_n - \mathbb{E}X) \longrightarrow N(0, \cdot)$

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 streaming data, ONLINE LEARNING, etc. (I will not cover this topic this semester.)

Q: Is data variety a problem?

Not really. The topic of multivariate analysis has existed since early 1900s.

Multivariate analysis. Given a sample $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ of random observations of dimension p, each $\mathbf{x_i} = [x_i^{(1)} \; x_i^{(2)} \; \cdots \; x_i^{(p)}]$ (or transposed version).

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

Observed Gaussian: Student's T-test, Fisher's test, ANOVA, all of which are *non-asymptotic* methods.

Non-Gaussian: (non-asymptotic) results are hard to obtain $\longrightarrow$ limiting theorems based on model statistics.

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

Classic MVA has a $p < 10$.

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 no longer applies**."

Classic paradigm:

- dimension p is small compared to the sample size n.
- asymptotic theory assumes n increases (very quickly to ∞) while dimension p remains fixed.
- At time t, we have all the data necessary for our analysis, i.e. the batch case.

No longer applies means:

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

Unique features of big data:

(Quick overview as I haven't presented notation yet). [Fan, Han, Liu; 2014] and references therein.

- **Heterogeneity:** With small data, data points from subpopulations are considered 'outliers'. With large data sets, subpopulations might be large. $\implies$ Mixtures of Gaussians?
- **Noise accumulation:** Errors accumulate when a decision 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,

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

  'endogeneity' means some features (predictors) correlate with the residual noise ϵ. That the residual noise ϵ is uncorrelated with all features is crucial. Called "Exogenous assumption" that $\mathbb{E}[\epsilon X_i] = 0$ for $i = 1, \ldots, p$. Easily violated in high-dimensions.

Aim of the course

Go from classic $\longrightarrow$ cutting-edge

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

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**

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pp} \end{bmatrix}, x_{ij} : \Omega \to \mathbb{R}(\text{or } \mathbb{C})$$

Q: What is a Random Matrix? [Diaconis 05]

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

$$X : \Omega \to \mathcal{S}, \mathcal{S} = \mathbb{R}, \mathbb{R}^2, \ldots$$

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 picking a matrix (with certain properties) at random with a certain probability.

E.g. Pick a random covariance matrix.

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

RMT Quantum mechanics 40's - 50's

- Energy levels of a system are described by eigenvalues of a hermitian operator on a Hilbert space.
- Computationally you can't work on infinite-dimension objects...
  - $\to$ discretization and truncation: keep only parts that are important to the problem under consideration.
  - $\to$ A finite but large random linear operator.
- Semicircular law for Gaussian (or Wigner) matrix [Wigner 1955, 1958] $\to$ [Arnold 1967, 1971] [Grenander 1963]
- Gaussian Wishart matrices (sample covariance matrices). [Marchenko/Pastur 1967] [Pastur 1972; 1973]. $\to$ Marchenko-Pastur law.
- Asymptotic theory of large sample covariance matrices. [Bai et al 1986] [Grenander, Silverstein 1977] [Johansson 1982] (?) ...
  Multivariate Fisher matrices (QR^{-1}), $Q \perp\!\!\!\perp R$ sample covariance matrices.
- Recently, 2$^{\text{nd}}$-order theory: CLT for linear spectral statistics, limit distribution spectral spacings, extreme eigenvalues.

Sample covariance matrices.

$\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_n$ sample of random observations with dimension p.

Population covariance matrix: $\sum = Cov(\mathbb{X}_i)$

Sample covariance matrix: $\mathcal{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbb{X}_i - \bar{\mathbb{X}})(\mathbb{X}_i - \bar{\mathbb{X}})^*$

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

Most results in MVA rely on $\mathcal{S}_n$: PCA, canonical correlation analysis, multivariate regression, one-sample or two-sample hypothesis testing, factor analysis.

$\Longrightarrow$ Understanding asymptotic properties of $\mathcal{S}_n$ extremely important in data analysis when p becomes large with respect to sample size n.

Generalized Variance and multiple correlation coefficient.

$\Longrightarrow$ overall measure of dispersion of the data, σ_i^2 measures $\mathbb{X}_i$, all variables together: generalized variance, "measure of scatter".

p becomes large $\Longrightarrow$ "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 $i^2 = -1$.

$$Re[a + ib] = a, \; Im[a + ib] = b$$

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

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

$$[A^*]_{ij} := \overline{A_{ji}} \text{ or } A^* := \left(\overline{A}\right)' = \overline{(A')}$$

The matrix $A = (a_{ij})$ is Hermitian if it is square with $a_j \in \mathbb{C}$ 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 $A^*A = AA^* = I$.

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

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

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 determinant 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:

- $(A+B)' = A' + B'$
- $(AB)' = B'A'$
- $\det(A') = \det(A)$
- $(A')^{-1} = (A^{-1})'$
- $tr(cA) = c \cdot tr(A)$
- $tr(A \pm B) = tr(A) \pm tr(B)$
- $tr(AB) = tr(BA)$
- $tr(B^{-1}AB) = tr(B)$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $tr(AA') = \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij}^2$
- $\det(AB) = \det(A)\det(B)$
- $\det(cA) = c^k \det(A)$