The Problem of Size

prof. dr Arno Siebes

Algorithmic Data Analysis Group
Department of Information and Computing Sciences
Universiteit Utrecht

Does Size Matter?

Volume

In the previous lecture we characterised Big Data by the three V's

- Volume, Velocity, and Variety

As we already discussed, Volume and Velocity have a lot in common. What we did not discuss is

- why is Volume a problem at all?

We will look at three aspects of this question:

- computational complexity (you are probably not surprised)
- the curse of dimensionality
- significance

A Small Network

The number of students enrolled in one of the department's programmes is in the order of 1500

- too big to know everyone
- but not dauntingly so

To support communication among the students and the staff

- one could build a simple CS-social network

From which we could directly compute fun facts and statistics like

- list all friends you have ($O(n)$)
- compute the average number of friends ($O(n^2)$)
- determine the friendliest student ($O(n^2)$)

and so on and so forth; all easily done on a bog standard computer

Facebook

Purely by coincidence, another social network has

- in the order of 1.5 billion (1.5×10^9) active users

Suppose that Facebook simply uses our (not very smart) implementation for the fun facts of the previous slide.

- If it takes us a millisecond to compute all your friends, it will take Facebook
  - one million milliseconds = 1000 seconds ≈ 15 minutes
- If it takes us a millisecond to determine the friendliest student, it will take Facebook
  - one million × one million milliseconds ≈ 1 million × 15 minutes ≈ 10,000 days ≈ 25 years

A billion is really a big number: even quadratic problems are a problem.

Preferably, algorithms should be $O(n \log n)$, or $O(n)$, or even better: sublinear. $O(n^3)$ is simply out of the question

The Curse of Dimensionality

While it may sound like the title of a comic book

- Tintin and the curse of dimensionality

it is actually the name of a serious problem for high dimensional data:

- high dimensional spaces are rather empty

And Big Data is often (very) high dimensional, e.g.,

- humans have in the order of 20,000 genes
- in novels one encounters 5000 - 10,000 distinct words

hence, it is important to be aware of this problem

But first: what does it mean that high dimensional space is empty?

d-Cubes

A little calculus shows that the volume of a d-dimensional cube C_d of width r is given by

$$V(C_d) = \int \cdots \int_{C_d} 1 \, dx_1 \dots dx_d = r^d$$

If we take a slightly smaller d-cube λC_d with width λr. we obviously have

$$V(\lambda C_d) = \lambda^d \times r^d = \lambda^d V(C_d)$$

Since for any $\lambda \in [0, 1)$ and for any $r \in \mathbb{R}$ we have that

$$\lim_{d \to \infty} \frac{V(\lambda C_d)}{V(C_d)} = \lim_{d \to \infty} \frac{\lambda^d V(C_d)}{V(C_d)} = \lim_{d \to \infty} \lambda^d = 0$$

we see that the higher d, the more of volume of C_d is concentrated in its outer skin of C_d: that is were the most points are.

d-Balls

Any first year calculus course teaches you that the volume of a d dimensional sphere S_d with radius r is given by

$$V(S_d) = \int \cdots \int_{S_d} 1 \, dx_1 \ldots dx_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} r^d$$

So again, for the d-ball λS_d we have

$$V(\lambda S_d) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \lambda^d r^d = \lambda^d V(S_d)$$

And, again, for any $\lambda \in [0,1)$ and for any $r \in \mathbb{R}$ we have that

$$\lim_{d \to \infty} \frac{V(\lambda S_d)}{V(S_d)} = \lim_{d \to \infty} \frac{\lambda^d V(S_d)}{V(S_d)} = \lim_{d \to \infty} \lambda^d = 0$$

Again, the volume is in an ever thinner outer layer.

d-Anything

This observation doesn't only hold for cubes and sphere. For, if you think about, it is obvious that for any (bounded) body B_d in $\mathbb{R}^d$ we have that

$$V(\lambda B_d) = \lambda^d V(B_d)$$

So, for all sorts and shapes we have that

> *the higher the dimension, the more of the volume is in an (ever thinner) outer layer*

In other words

> *In high dimensional spaces, points are far apart*

Yet Another Illustration

Another way to see this is to consider a d-cube of width $2r$ and its inscribed d-ball with radius r:

$$\lim_{d\to\infty} \frac{\left(\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} r^d\right)}{(2r)^d} = \lim_{d\to\infty} \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)2^d} = 0$$

If we have a data point and look at the other points within a given distance

> *we'll find fewer and fewer the higher d is.*

That is, again we see that

> *in high dimensional spaces, points are far apart*

In fact, under mild assumptions[1] all points are equally far apart! That is, you are searching for the data point nearest to your query point: and the all are equally qualified.

[1] When is "Nearest Neighbor" Meaningful, Beyer et al, ICDT'99

So, Why is this Bad? Similarity

The assumption underlying many techniques is that

- similar people behave similarly

For example,

- if you are similar to (a lot of) people who repayed their loan, you will probably repay
- if (many) people similar to you liked Harry Potter books, you'll probably like Harry Potter books

It is a reasonable assumption

- would we be able to learn if it doesn't hold at all?

and it works pretty well in practice. But what if

- no one resembles you very much?
- or everyone resembles you equally much?

in such cases it isn't a very useful assumption

Why is it Bad? Lack of Data

Remember, we try to learn the data distribution. If we have d dimensions/attributes/features... and each can take on v different values, then we have

$$v^d$$

different entries in our contingency table. To give a reasonable probability estimate, you'll need a few observations for each cell. However

- v^d is quickly a vast number, overwhelming the number of Facebook users easily.

After all,

$$2^{30} > 10^9$$

and 30 is not really high dimensional, is it? And 2^{40} is way bigger than 10^9

So, we talk about Big Data, but it seems we have a lack of data!

Are We Doomed?

The curse of dimensionality seems to make the analysis of Big Data impossible:

- we have far too few data points
- and the data points we have do not resemble each other very much

However, life is not that bad:

> *data is often not as high-dimensional as it seems*

After all, we expect *structure*

- and structure is a great dimensionality reducer

One should, however, be aware of the problem and techniques such as *feature selection* and *regularization* are very important in practice.

Significance

The first two consequences of "Big" we discussed

- computational complexity and
- the curse of dimensionality

are obviously negative: "Big" makes our life a lot harder.

For the third, significance, this may seem different

- "Big" makes everything significant

However, that is not as nice as you might think. Before we discuss the downsides, let us first discuss

- statistics and their differences
- what we mean by significance
- and the influence of "Big" on this

Statistic

A statistic is simply a, or even *the*, property of the population we are interested in. Often this is an *aggregate* such as the *mean* weight.

If we would have access to the whole population – if we knew the distribution $\mathcal{D}$ – we would talk about a parameter rather than a statistic. We, however, have only a sample – D – from which we compute the statistic to *estimate* the parameter.

And, the natural question is:

how good is our estimate?

Slightly more formal, how big is

$$\|\beta - \hat{\beta}\|?$$

Sampling Distribution

The problem of using a sample to estimate a parameter is that we may be unlucky

- to estimate the average height of Dutch men, we happen to pick a Basketball team

The statistic itself has a distribution over all possible samples

- each sample yields its own estimate

This distribution is known as the *sampling distribution*

The question how good our estimate is depends on the sampling distribution, There are well-known bounds

- without assumptions on the data distribution
- but also for given distributions (obviously tighter)

Before we discuss such bounds, we first recall the definitions of Expectation and Variance

Expectation

For a random variable X, the *expectation* is given by:

$$\mathbb{E}(X) = \sum_{\Omega} x \times \mathbb{P}(X = x)$$

More general, for a function $f : \Omega \to \mathbb{R}$ we have

$$\mathbb{E}(f(X)) = \sum_{\Omega} f(x) \times \mathbb{P}(X = x)$$

Expectation is a linear operation:

1. $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
2. $\mathbb{E}(cX) = c\mathbb{E}(X)$

Expectation of a Sample

Let X_i be independent identically distributed (i.i.d) random variables

- e.g., the X_i are independent samples of the random variable X

Consider the new random variable

$$\frac{1}{m}\sum_{i=1}^{m} X_i$$

Then

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m} X_i\right) &= \frac{1}{m}\mathbb{E}\left(\sum_{i=1}^{m} X_i\right) \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(X_i) \\
&= \frac{m}{m}\mathbb{E}(X) = \mathbb{E}(X)
\end{aligned}$$

Conditional Expectations

Like conditional probabilities there are conditional expectations. Let $F \subseteq \Omega$ be an event, then

$$\mathbb{E}(X \mid F) = \sum_{\Omega} x \times \mathbb{P}(X = x \mid F)$$

If a set of events $\{F_1, \ldots, F_n\}$ is partition of Ω, i.e.,

- $\forall i, j \in \{1, \ldots, n\} : i \neq j \Rightarrow F_i \cap F_j = \emptyset$
- $\bigcup_{i \in \{1, \ldots, n\}} F_i = \Omega$

then

$$\mathbb{E}(X) = \sum_{i} \mathbb{P}(F_i)\mathbb{E}(X \mid F_i)$$

that is, the unconditional expectation is the weighted average of the conditional expectations

Variance

The *variance* of a random variable is defined by

$$\sigma^2(X) = Var(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$$

The *standard deviation* is the square root of the variance

$$\sigma(X) = \sqrt{Var(X)} = \sqrt{\mathbb{E}((X - \mathbb{E}(X))^2)}$$

Some simple, but useful, properties of the variance are:

1. $Var(X) \geq 0$
2. for $a, b \in \mathbb{R}$, $Var(aX + b) = a^2 Var(X)$
3. $Var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$
4. $Var(X) \leq \mathbb{E}(X^2)$

Variance of a Sample

Let X_i be independent identically distributed (i.i.d) random variables

- e.g., the X_i are independent samples of the random variable X

Consider again the random variable

$$\frac{1}{m}\sum_{i=1}^{m} X_i$$

Then

$$\begin{aligned} Var\left(\frac{1}{m}\sum_{i=1}^{m} X_i\right) &= \frac{1}{m^2} Var\left(\sum_{i=1}^{m} X_i\right) \\ &= \frac{m}{m^2} Var(X) = \frac{Var(X)}{m} \end{aligned}$$

The larger the number of samples, the smaller the variance

Covariance

If we have two random variables X and Y, their *covariance* is defined by

$$Cov(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$$

or, equivalently, by

$$Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Which immediately tells us that if X and Y are independent, then their covariance $Cov(X, Y) = 0$. Note that the reverse is *not* true.

Moreover,

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

Correlation

Although we will not use it today, it would feel odd to recall covariance but not its normalised version known as *correlation.* If both $Var(X)$ and $Var(Y)$ are finite, their correlation is given by:

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

From this definition it is easy to verify that

$$-1 \leq Corr(X, Y) \leq 1$$

- If there is a linear relation between X and Y, i.e., $Y = aX$, then $|Corr(X, Y)| = 1$
- If X and Y are independent then $Corr(X, Y) = 0$

Note that again $Corr(X, Y) = 0$ does *not* imply independence

- in fact Y may be completely determined by X

For that reason, *mutual information* may be a better estimate of the relationship between X and Y.

Markov's Inequality

With Expectation and Variance knowledge refreshed, let us return to the quality of our estimates. The first question is:

> *What is the probability that the value of a random variable X is far from its expectation?*

This question is answered by Markov's inequality:

For a non-negative random variable $X : \Omega \to \mathbb{R}$, i.e., $X(e) \geq 0$, and positive real number a:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

Clearly, this isn't a very strong bound, e.g.,

- the probability that $X \geq \mathbb{E}(X)$ is bounded by 1
- the probability that $X \geq a\mathbb{E}(X)$ is bounded by $\frac{1}{a}$

but it does hold for all possible distributions!

Proof

Let $Y = \{e \in \Omega \mid X(e) \geq a\}$, then

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{\Omega} X(e)\mathbb{P}(e) \\
&= \sum_{Y} X(e)\mathbb{P}(e) + \sum_{\Omega \setminus Y} X(e)\mathbb{P}(e) \\
&\geq \sum_{Y} X(e)\mathbb{P}(e) \quad (\forall s : X(e)\mathbb{P}(e) \geq 0) \\
&\geq \sum_{Y} a\mathbb{P}(e) \quad (\forall e \in Y : X(e) \geq a) \\
&= a \sum_{Y} \mathbb{P}(e) = a\mathbb{P}(Y)
\end{aligned}$$

That is, $\mathbb{E}(X) \geq a\mathbb{P}(X \geq a)$ and we are done.

Chebyshev's Inequality

Markov's inequality doesn't refer to the variance. His advisor Chebyshev has an inequality that does:

Let $X : \Omega \to \mathbb{R}$ be a random variable and let $a > 0$ be a real number then:

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{Var(X)}{a^2}$$

The proof is easy, use the random variable $(X - \mathbb{E}(X))^2$ and plug it into Markov's inequality:

$$\begin{aligned} \mathbb{P}(|X - \mathbb{E}(X)| \geq a) &= \mathbb{P}((X - \mathbb{E}(X))^2 \geq a^2) \\ &\leq \frac{\mathbb{E}((X - \mathbb{E}(X))^2)}{a^2} = \frac{Var(X)}{a^2} \end{aligned}$$

Chebyshev on a Sample

Let X_i be independent identically distributed (i.i.d) random variables

- e.g., the X_i are independent samples of the random variable X

such that $Var(X_i) < 1$ and denote $\mathbb{E}(X_i) = \mu$ Then for any $\delta \in (0, 1)$ we have that

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| \leq \sqrt{\frac{1}{\delta m}}\right) \geq 1 - \delta$$

That is, with sample size of 100, we are already for 99% sure that our sample average is within a distance of 1 of the distribution's mean.

More in general, the difference is bounded by the square root of the sample size.

Proof

Consider the random variable:

$$\frac{1}{m}\sum_{i=1}^{m} X_i$$

and recall that

- $\mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m} X_i\right) = \mathbb{E}(X) = \mu$
- $Var\left(\frac{1}{m}\sum_{i=1}^{m} X_i\right) = \frac{Var(X)}{m}$

Plug it into Chebyshev's inequality and we get:

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| \geq a\right) \leq \frac{Var(X)}{ma^2} \leq \frac{1}{ma^2}$$

Set $\delta = \frac{1}{ma^2}$, i.e., $a = \sqrt{\frac{1}{\delta m}}$ and we are done.

Chernoff's Bounds

If we know more about the X_i we can derives tighter bounds.

Let $X = \sum_{i=1}^{n} X_i$ where $\mathbb{P}(X_i) = p_i, \mathbb{P}(X_i = 0) = 1 - p_i$ and the X_i are independent. Let $\mu = \mathbb{E}(X) = \sum_1^n p_i$, then

1. $\forall \delta > 0 : \mathbb{P}(X \geq (1 + \delta)\mu) \leq e^{-\frac{\delta^2}{2+\delta}\mu}$
2. $0 < \delta < 1 : \mathbb{P}(X \leq (1 - \delta)\mu) \leq e^{-\frac{\delta^2}{2}\mu}$
3. Hence, $0 < \delta < 1 : \mathbb{P}(|X - \mu|) \geq \delta\mu) \leq 2e^{-\frac{\mu\delta^2}{3}}$

Note we are not going to prove these bounds. Next, note that if all the p_i are the same we talk about Bernoulli trials, otherwise it is known as Poisson trials.

Example: Coin Tosses

Let the X_i represent tosses of a fair coin with $p_i = 0.5$ Denote by S_n the number of heads in n tosses, i.e., $S_n = \sum_1^n X_i$ and $\mathbb{E}(S_n) = \frac{n}{2}$. Then.

Chebyshev:
$\mathbb{P}(|S_n/n - 1/2| \geq \epsilon) \leq \frac{1}{4n\epsilon^2}$. If we choose $\epsilon = 1/4$, we get:

$$\mathbb{P}(|S_n/n - 1/2| \geq 1/4) \leq 4/n$$

Chernoff:
$\mathbb{P}(|S_n - n/2| \geq \delta n/2) \leq 2e^{-n\delta^2/6}$. Choose $\delta = 1/2$ and we get

$$\mathbb{P}(|S_n/n - 1/2| \geq 1/4) \leq 2e^{-n/24}$$

That is, Chernoff is massively smaller that Chebyshev: knowing the distribution gives you a much tighter bound

Hoeffding's Inequality

The concentration measure that we will use over and over again is by Hoeffding.

Let $Z_1, \ldots, Z_m$ be a sequence of i.i.d. random variables and let $\bar{Z} = \frac{1}{m}\sum_{i=1}^{m} Z_i$. Furthermore, Let $\mathbb{E}(\bar{Z}) = \mu$ and assume that $\mathbb{P}[a \leq Z_i \leq b] = 1$, for every i. Then, for any $\epsilon > 0$, we have

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon\right] \leq 2e^{-\left(\frac{2m\epsilon^2}{(b-a)^2}\right)}$$

We will not prove it right now, but later in this course we will prove a slightly stronger result (from which Hoeffding easily follows)

So, What?

We introduced a number of concentration measures and you might be thinking:

> *so what?*

The pragmatic reason is that we will use (some of these) later in the course. The more important reason is that all these measures tell us the same thing:

> *the larger the sample, the closer our statistic is probably to the true parameter*

While this is intuitively obvious,

- these bounds tell you how close you can expect to be
- and how fast this scales with size

In fact, in a Big Data world, they tell us

- we can expect that all our estimates will be pretty accurate

Blessing and Curse

Clearly, this is good news

- or, is it?

Well, the answer is yes and no

Yes obviously it is good that we have an accurate view of the world through our Big Data lens.

No because it will make even tiny differences appear significant

To understand the latter we need to dive into statistical tests. But it is already useful to consider the following (bogus) fact

- young men from Utrecht are (on average) significantly larger than young men from Houten with a difference of 0.1 mm

is this useful or not? Is it the type of knowledge you hope Big Data will bring us?

Statistical Testing

Suppose that we have taken a sample from the young men in Utrecht we measured them, and perhaps even computed their average height. Then we meet a new young man, can we say something about

- how *likely* he is to live in Utrecht given his height?

Or we have sampled young men from both Utrecht and Houten and computed the average height for both samples, Can we say

- whether or not both populations have the same average height or not?

Or, many similar questions

This is the realm of *statistical testing* and it depends very much on the *sampling distribution* we already met.

Question 1

We can turn our measurements of the heights of our sample of young Utrecht men into a nice histogram

- if the height of our new acquaintance is somewhere smack in the middle of this histogram, we have no reason to believe that he is *not* living in Utrecht
- if he is, however, far taller than anyone in our histogram, we would not be surprised to learn that he actually comes from Brobdignag.

The crucial number people look at is

$$\mathbb{P}(X \geq l_{new})$$

also known as a *p value* The important point (for now) is that

- you look at a histogram and decide from there whether or not something is likely or not.

Question 2

When we sample from a population to estimate a parameter by computing a statistic

- we know that this statistic is governed by a sampling distribution

That is, if we have two different samples, the statistic will be different for the two samples.

Now we have two samples and, thus, two statistics and we wonder whether these two samples come from

- one and the same population (there is no difference between Utrecht and Houten)
- or from two different populations (young men from Utrecht are (on average) taller (or smaller) then young men from Houten

How do we decide between these options?

Given the Sampling Distribution

Assume that you know the sampling distribution of, say, the height of young men from Utrecht, that is you have

- a histogram of all average heights of all possible samples

and you notice that the average height of the sample of young men from Houten is smaller than 99% of the average height of all possible samples of young men from Utrecht

- than it seems reasonably safe to conclude that young men from Houten seem (on average) smaller than young men from Utrecht

In fact, you could say

- that 99% of Utrecht samples would have a larger average height
- and hence you are 99% sure that the two populations are different.

Using Both Sampling Distributions

If we have both sampling distribution histograms, life is even better

- you can estimate a p value for the Houten sample to be from the Utrecht sampling distribution
- and vice versa

If you are sure that your sample is either from Houten or from Utrecht

- which is very much true in our example

The *Bayes optimal* decision is to chose for the population with the largest probability

What do the two p values you estimated tell you about this choice?

How to get this Distribution

Our discussion on the preceding slides assumed that we have access to the sampling distribution. In general we only have *one* sample and no easy (affordable) way to get many more

- so it seems that we cannot use these ideas in practice

Fortunately, that isn't true. There are two ways out

- if we have reasons to believe that a statistic, like the height, follows a known distribution – like a *Gaussian* a.k.a. *Normal* distribution – we can simply compute the p value
  - this is the assumption that underlies much of the statistical testing theory
- in all other cases, we can pull ourselves out of the problems by our bootstraps
  - yes, named after one of the tales of the (in)famous Baron (von) Münchhausen

The Bootstrap

To go from one sample to many we use *resampling*.

- given a data set D
- create a bootstrap sample D'
  - sample a random element from D
  - exactly $|D|$ times (with replacement)
- and create many such equally sized samples

If we compute our statistic on each of these bootstrap samples

- we get a distribution that approximates the sampling distribution.

Why Does the Bootstrap Work?

The intuition is not that difficult

- If you have a very, very large sample.
  - sampling from that sample will be very similar to sampling from the distribution
- For smaller samples (data sets)
  - note that each “object” in your sample “represents” multiple objects from the distribution
  - by sampling with replacement you simulate the possibility that you would sample more objects with the same characteristics from the distribution

The proof that it works is elegant

- but uses some advanced maths for which unfortunately do not have time

Large Means Narrow

The concentration measures we discussed today tell us that

> *large samples have a narrow sampling distribution*

That is,

- almost all data is close to the mean

Since p values are concerned with the probability that you are “this far” from the mean

- almost everything will have a small p value

That is,

> *the smallest difference will be statistically significant*

Which is not the same as significant in the sense of (practically) useful.

Spurious Correlations

There is another reason why this is bad:

> *Big Data means many spurious correlations*

In fact, using, e.g.,

- Ramsey Theory, or
- Ergodic Theory, or
- Algorithmic Information Theory

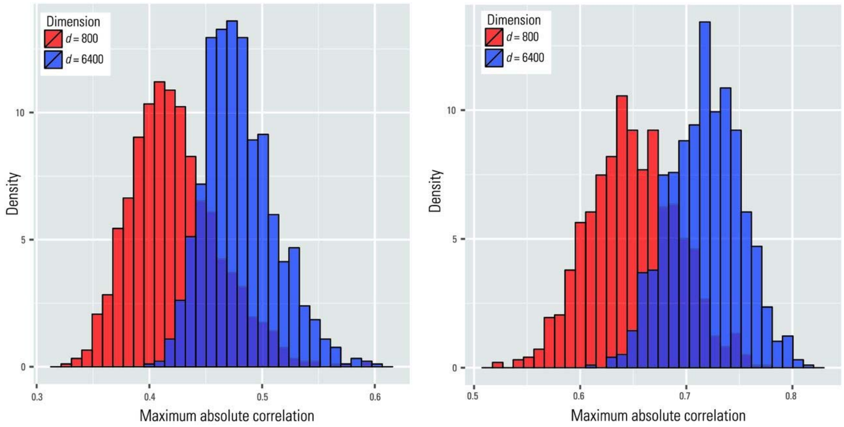
one can actually prove that there will be correlations in big data, even if it is completely random!
We will just look at an experiment of Jianqing Fan et al (Challenges of Big Data analysis). Generate d independent $N(0, 1)$ samples – correlations are expected to be 0 – and look at

- the largest correlation of an X_j with X_1 (left) and
- the maximal correlation of a weighted linear sum of 4 variables (regression) and X_1 (right)

Spurious Correlations, The Picture

11/14/2016

F2.large.jpg (1280×646)



http://nsr.oxfordjournals.org/content/1/2/293/F2.large.jpg

1/1

Moreover, Multiple Testing

There is a another testing problem, when we have Big Data

- we will often check very many hypotheses

And if you test 20 random(!) hypotheses with a p value of 0.05

- you will on average have 1 significant result
  - while being completely random

If you do many tests, you have correct your p values to be (more) sure of seeing real effects. The simplest one is the

- Bonferonni correction

The rule is

- if you test m hypotheses for a significance of α
- you claim success for those whose p value is $\leq \alpha / m$

Note that Bonferroni is a rather conservative test

- significant results may be discarded as not significant

There are alternatives like the Holm procedure

Conclusions

Don't get me wrong

- Big Data is good

It allows us to learn many things that were previously

- unattainable or (at least) hard

However, Big Data comes with its own problems, due to

- complexity, emptiness, and significance

Hence, we have to be careful

- in what we want to learn and how

Fortunately, as we will see in this course

- sampling is a good way to alleviate some of our problems