Big Data

High-Dimensional Data Sets

Cor Kraaikamp

. . . or how our intuition fails us

High-Dimensional Data Sets

In a High-Dimensional Data Set the number of "data points" n isn't necessarily high, but the number of components (parameters/covariates) p per data point is.

We will see that a consequence of this is, that such datasets behave quite different from what we expect/what our intuition tells us.

In this first lecture on Big Datasets and Stochastics I will try to make this plausible.

But first some examples of such datasets:

Examples of High-Dimensional Data Sets-continued

Consumer preferrence data

Companies like Amazon or bol.com keep track as your behavior as a customer. They use this data to make you personal offers which are tailored to your lifestyle.

Crowdserving

Companies like Google and Facebook keep track of your behavior on the web. This is interesting data for their customers who put adds on Google or Facebook. These customers can then target you individually as a user of Google or Facebook.

This browsing information is also very interesting to various government agencies, who like to know in detail which pages you visit, and which messages you leave behind on the web.

Clearly this list is far from exhaustive. Other categories you could think of are Business data and Financial data, Military data, ...

Curse of dimensionality

High-dimensionality impacts on statistics in various ways. We mention four of these briefly, and then will return to some of them in more detail.

First (and very importantly), high-dimensional spaces are vast/enormous and due to this (data)points are isolated in the immensity of their high-dimensional space.

Another impact is, that small fluctuations in (very) many directions may cause a large global fluctuation.

A third reason is, that the accumulation of rare events may itself be not rare.

Finally, numerical computations and optimizations in high-dimensional spaces can be overly "expensive" (in time, power, computer resources and -capacity).

High-Dimensional Datasets are vast

Suppose we want to explain a responce variable $Y \in \mathbb{R}$ by p real variables $X_1, X_2, \ldots, X_p$.

For sake of simplicity we assume that these X_i are i.i.d. uniformly $[0,1]$-distributed random variables.

In this case the p-dimensional random variable

$$X = (X_1, X_2, \ldots, X_p)$$

is uniformly distributed on the hypercube $[0,1]^p$.

High-Dimensional Datasets are vast

The most simple approach is the k-*nearest neighbor estimator*, where $f(x)$ is estimated by the mean of the Y_i associated to the k points $X^{(i)}$, which are nearest from x.

A little bit more sophisticated beyond this would be a weighted average with weights that are a decreasing function of the distance $||X^{(i)} - x||$ (think of kernel smoothing). In both cases the idea is to use a **local** average of the data.

This works well in low-dimensional data, but **not** in high-dimensional data!

High-Dimensional Datasets are vast

To get some "feeling" for these observations, assume that U and U' are two independent, uniformly distributed random variables on $[0,1]$.

Then the mean square distance between $X^{(i)}$ and $X^{(j)}$ (of course with $i \neq j$) is:

$$\begin{aligned}
\mathrm{E}\left(||X^{(i)} - X^{(j)}||^2\right) &= \mathrm{E}\left(\left(\sqrt{(X_1^{(i)} - X_1^{(j)})^2 + \cdots + (X_p^{(i)} - X_p^{(j)})^2}\right)^2\right) \\
&= \mathrm{E}\left(\sum_{k=1}^{p} (X^{(i)} - X^{(j)})^2\right) \\
&= \sum_{k=1}^{p} \mathrm{E}\left((X_k^{(i)} - X_k^{(j)})^2\right).
\end{aligned}$$

High-Dimensional Datasets are vast

So the standard deviation of this mean square distance is:

$$\text{sdev}\left(||X^{(i)} - X^{(j)}||^2\right) \approx 0.2\sqrt{p}.$$

Thus we see that the "typical" mean square distance between two sample points sampled uniformly in $[0,1]^p$ grows linearly with p, while the scaled deviation

$$\frac{\text{sdev}\left(||X^{(i)} - X^{(j)}||^2\right)}{\text{E}\left(||X^{(i)} - X^{(j)}||^2\right)} \approx \frac{0.2p}{p/6} = \frac{1.2}{\sqrt{p}},$$

shrinks like $1/\sqrt{p}$.

Again a confirmation that the concept of "local" gets lost when the dimension p grows large.

How many observations do we need?

One can show (again this is an *exercise* for Friday) that the volume $V_p(r)$ of a closed p-dimensional ball of radius $r > 0$ is given by:

$$V_p(r) = \frac{\pi^{p/2}}{\Gamma(p/2+1)} \approx \left(\frac{2\pi \mathrm{e} r^2}{p}\right)^{p/2} \frac{1}{\sqrt{p\pi}}, \qquad \text{for large } p \qquad (1)$$

where Γ is the famous “Gamma-function,” defined by:

$$\Gamma(x) = \int_0^\infty t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t, \qquad \text{for } x > 0.$$

How many observations do we need?

As a consequence,

$$1 = V_p\left([0,1]^p\right) \leq \sum_{i=1}^{n} V_p\left(B_p(x^{(i)},1)\right),$$

and from (1) it follows that:

$$1 \leq n \cdot \frac{\pi^{p/2}}{\Gamma(p/2+1)},$$

i.e.

$$n \geq \frac{\Gamma(p/2+1)}{\pi^{p/2}} \approx \left(\frac{p}{2\pi \mathrm{e}}\right)^{p/2} \sqrt{p\pi} \qquad \text{(for large } p\text{)}.$$

So we see that n should grow more than exponentially fast with p.

Fluctuation accumulate

Suppose we have some scalar $\theta_1 \in \mathbb{R}$, and that we want to evaluate some function $F(\theta_1)$ of θ_1. Due to noise we only have access to

$$X_1 = \theta_1 + \varepsilon_1,$$

where ε_1 is a random value with $\mathrm{E}(\varepsilon_1) = 0$ and $\mathrm{Var}(\varepsilon_1) = \sigma^2 (> 0)$.

If the function F is 1-Lipschitz, then we have that:

$$\begin{aligned} |F(X_1) - F(\theta_1)| &= |F(\theta_1 + \varepsilon_1) - F(\theta_1)| \\ &\leq 1 \cdot |(\theta_1 + \varepsilon_1) - \theta_1| \\ &= |\varepsilon_1|. \end{aligned}$$

But then we have (since $y = x^2$ is a monotonically increasing function on $\mathbb{R}^+$):

$$|F(X_1) - F(\theta_1)|^2 \leq |\varepsilon_1|^2$$

and from this we find, that:

$$\mathrm{E}\left(|F(X_1) - F(\theta_1)|^2\right) \leq \mathrm{E}\left(|\varepsilon_1|^2\right) = \sigma^2.$$

So in one dimension (i.e. $p = 1$) things are pretty nice!

Fluctuation accumulate

Now one might argue that $p\sigma^2$ is just an upper bound for

$$\mathrm{E}\left(||F(X_1,\ldots,X_p) - F(\theta_1,\theta_2,\ldots,\theta_p)||^2\right);$$

the actual value of this expected value might be much smaller!

However, if $||F(x+h) - F(x)|| \geq c \cdot ||h||$ for some $c > 0$, then:

$$||F(X_1,\ldots,X_p) - F(\theta_1,\ldots,\theta_p)||^2 \geq c \cdot ||(\varepsilon_1,\ldots,\varepsilon_p)||^2,$$

but then the mean square error error

$$\mathrm{E}\left(||F(X_1,\ldots,X_p) - F(\theta_1,\ldots,\theta_p)||^2\right)$$

scales like $p\sigma^2$.

An example where this situation might arise is the linear regression model with high-dimensional covariates.

High-Dimensional Linear Regression

A classical estimator of $\beta^\star$ is the least squares estimator $\hat{\beta}$, defined by

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} ||Y - \mathbf{X}\beta||, \tag{2}$$

which is uniquely defined if the rank of $\mathbf{X}$ is p. For simplicity we focus on this case. Then it is well-known from Linear Algebra that the solution of the minimization problem (2) is given by:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y.$$

One can show that:

$$\mathrm{E}\left(||\hat{\beta} - \beta^\star||^2\right) = \mathrm{E}\left(||(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon||^2\right) = \mathrm{Tr}\left((\mathbf{X}^T\mathbf{X})^{-1}\right)\sigma^2.$$

Computational Complexity

Another burden arises in high-dimensional settings: numerical computations can become very intensive and easily exceed the available computational (and memory) resources.

We just saw in our regression model-example that the mean square error $||\hat{\beta} - \beta^{\star}||^2$ in the linear regression model

$$y = \sum_{j=1}^{p} \beta_j^{\star} x_j + \varepsilon$$

typically scales linearly with p. Of course, it is unlikely that all the covariates x_j influence the responce y.

Tails of high-dimensional Gaussian distributions are thin but contain most of the mass

Gaussian distributions are known to have very thin tails. In fact, the density

$$g_p(x) = \frac{1}{(\sqrt{2\pi})^p} \cdot \mathrm{e}^{-\frac{1}{2}||x||}$$

of a standard Gaussian distribution (i.e. a $N(0, I_p)$-distributed random variable) in $\mathbb{R}^p$ decreases exponentially fast with the square norm of x.

Yet ... when p is large, most of the mass of the standard Gaussian distribution lies in its tails!!

How can we see this?

Tails of high-dimensional Gaussian distributions are thin but contain most of the mass

Recall the Markov Inequality:

Markov Inequality

For any non-decreasing positive function $\psi : \mathbb{R} \to \mathbb{R}^+$ and any real-valued random variable X, we have

$$\mathrm{P}\left(X \geq t\right) \leq \frac{1}{\psi(t)} \mathrm{E}\left(\psi(X)\right), \qquad \text{for all } t \in \mathbb{R}.$$

In particular, for any $\lambda > 0$ we have

$$\mathrm{P}\left(X \geq t\right) \leq \mathrm{e}^{-\lambda t} \mathrm{E}\left(\mathrm{e}^{\lambda X}\right), \qquad \text{for all } t \in \mathbb{R}.$$

Tails of high-dimensional Gaussian distributions are thin but contain most of the mass

In fact, we shall use an "easier" version of this inequality:

Markov Inequality

If X is any nonnegative integrable random variable and $a > 0$, then

$$\mathrm{P}\left(X \geq a\right) \leq \frac{\mathrm{E}\left(X\right)}{a}.$$

From the "easy version" of Markov's Inequality we find:

$$\begin{aligned}
\mathrm{P}\left(X \in B_{p,\delta}\right) &= \mathrm{P}\left(\mathrm{e}^{-||X||^2/2} \geq \delta\right) \\
&\leq \frac{1}{\delta}\mathrm{E}\left(\mathrm{e}^{-||X||^2/2}\right) \\
&= \frac{1}{\delta}\int_{x\in\mathbb{R}^p} \mathrm{e}^{-||x||^2}\,\frac{\mathrm{d}x}{(2\pi)^{p/2}} \\
&= \frac{1}{\delta 2^{p/2}}.
\end{aligned}$$

More counter-intuitive phenomena ... and the way out

In fact, in Chapter 1 of *Introduction to High-Dimensional Statistics* by Christophe Giraud the previous examples plus a few more are given of *counter-intuitive phenomena*. Clearly I cannot deal with them all ... but the message should be clear by now! One cannot simply rely on one's intuition (from low-dimensional data) in a high-dimensional setting!

The thing which at first seemed to be a blessing now looks like a curse! In fact, the situation might appear hopeless to you now.

Fortunately, high-dimensional data are often much more low-dimensional than they at first appear to be. Usually they are not "spread out uniformly" across $\mathbb{R}^p$, but rather clustered around lower-dimensional structures. These structures are due to low complexity of the systems producing the data. Giraud lists various examples:

A Paradigm Shift

In fact, in his introductory chapter Giraud claims that a *Paradigm Shift* is needed. In his view, classical statistics provide a very rich theory for analysing data with the following characteristics:

- a small number p of parameters
- a large number n of observations

As we can see from his examples (some of which I just discussed), in many fields data have very different characteristics:

- a huge number p of parameters
- a sample size n which is either roughly the size of p, or sometimes much smaller than p.

A Paradigm Shift

One of the possible approaches (the one Giraud advocates) is to treat n and p as they are and provide a non-asymptotic analysis of the estimators, which holds for any n and p. Giraud warns that the drawback of such a method (above the classical asymptotic analysis in whch n tends to infinity) is that it is much more involved; usually one needs much more elaborate arguments in order to provide precise enough results.

Mathematics of High-Dimensional Statistics

But then it follows from the CLT that for a L-Lipschitz function f and i.i.d. random variables $X_1, X_2, \ldots, X_n$ with finite variance σ^2 we have that:

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_1) - \mathrm{E}(f(X_1)) \geq \frac{L\sigma}{\sqrt{n}}x\right) \leq \mathrm{P}\left(Z \geq x\right) \leq \mathrm{e}^{-x^2/2},$$

for $x > 0$.

Concentration inequalities provide some non-asymptotic versions of such results.

Gaussian Concentration Inequality

Assume that $F : \mathbb{R}^d \to \mathbb{R}$ is 1-Lipschitz and that Z has a Gaussian $N(0, \sigma^2 I_d)$ distribution. Then there exists a variable ξ with exponential distribution with parameter 1, such that

$$F(Z) \leq \mathrm{E}(F(Z)) + \sigma\sqrt{2\xi}.$$

Mathematics of High-Dimensional Statistics

According to the Gaussian Concentration inequality, we then have for $x > 0$ and $n \in \mathbb{N}$

$$\mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_1) - \mathrm{E}(f(X_1)) \geq \frac{L\sigma}{\sqrt{n}}x\right) \leq \mathrm{P}\left(\sqrt{2\xi} \geq x\right) = \mathrm{e}^{-x^2/2},$$

which can be viewed as an non-asymptotic version of (**??**).