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.

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.

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.

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

Biotech data

DNA micro-arrays contain the information on tens of thousands of genes of single individuals. Usually the number of individuals n in such a dataset is small.

Examples of High-Dimensional Data Sets

Biotech data

DNA micro-arrays contain the information on tens of thousands of genes of single individuals. Usually the number of individuals n in such a dataset is small.

Clearly, for the development of future medicine it is important to understand the relation between these p genes.

Examples of High-Dimensional Data Sets

Biotech data

DNA micro-arrays contain the information on tens of thousands of genes of single individuals. Usually the number of individuals n in such a dataset is small.

Clearly, for the development of future medicine it is important to understand the relation between these p genes.

Images (and videos)

Large datasets of images are collected almost continuously around the world;

Examples of High-Dimensional Data Sets

Biotech data

DNA micro-arrays contain the information on tens of thousands of genes of single individuals. Usually the number of individuals n in such a dataset is small.

Clearly, for the development of future medicine it is important to understand the relation between these p genes.

Images (and videos)

Large datasets of images are collected almost continuously around the world; Facebook, video surveillance images, medical images, etc.

Examples of High-Dimensional Data Sets

Biotech data

DNA micro-arrays contain the information on tens of thousands of genes of single individuals. Usually the number of individuals n in such a dataset is small.

Clearly, for the development of future medicine it is important to understand the relation between these p genes.

Images (and videos)

Large datasets of images are collected almost continuously around the world; Facebook, video surveillance images, medical images, etc. Each image (or video) consists of massive amounts of pixels.

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.

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.

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.

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.

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.

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.

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,

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, ...

Blessing?

Being able to “see” simultaneously thousands (or more) variables on each individual (datapoint) seems a *great & wonderful thing*.

Blessing?

Being able to "see" simultaneously thousands (or more) variables on each individual (datapoint) seems a *great & wonderful thing*.

After all, potentially we may scan/gather information on every variable that may influence the phenomenon under study.

Blessing?

Being able to "see" simultaneously thousands (or more) variables on each individual (datapoint) seems a *great & wonderful thing*.

After all, potentially we may scan/gather information on every variable that may influence the phenomenon under study.

This sounds like great news . . .

Blessing?

Being able to "see" simultaneously thousands (or more) variables on each individual (datapoint) seems a *great & wonderful thing*.

After all, potentially we may scan/gather information on every variable that may influence the phenomenon under study.

This sounds like great news ... unfortunately, the (statistical) reality clashes harshly with this optimistic point of view;

Blessing?

Being able to "see" simultaneously thousands (or more) variables on each individual (datapoint) seems a *great & wonderful thing*.

After all, potentially we may scan/gather information on every variable that may influence the phenomenon under study.

This sounds like great news ... unfortunately, the (statistical) reality clashes harshly with this optimistic point of view; separating the data/signal from the noise is *in general* almost impossible in high-dimensional data due to the so-called "curse of dimensionality."

Curse of dimensionality

High-dimensionality impacts on statistics in various ways.

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.

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

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.

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.

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.

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).

Goals of this class

In this introductory class I want to address some of these four aspects by rather simple examples,

Goals of this class

In this introductory class I want to address some of these four aspects by rather simple examples, just to give you a "feel" of what is going on.

Goals of this class

In this introductory class I want to address some of these four aspects by rather simple examples, just to give you a “feel” of what is going on.

All these example are from Christophe Giraud’s book *Introduction to High-Dimensional Statistics*.

Goals of this class

In this introductory class I want to address some of these four aspects by rather simple examples, just to give you a "feel" of what is going on.

All these example are from Christophe Giraud's book *Introduction to High-Dimensional Statistics*.

In fact, the only thing I would like to achieve today is to make you aware that your intuition concerning the statistics of high-dimensional data is

Goals of this class

In this introductory class I want to address some of these four aspects by rather simple examples, just to give you a “feel” of what is going on.

All these example are from Christophe Giraud’s book *Introduction to High-Dimensional Statistics*.

In fact, the only thing I would like to achieve today is to make you aware that your intuition concerning the statistics of high-dimensional data is probably *wrong*.

High-Dimensional Datasets are vast

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

High-Dimensional Datasets are vast

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

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

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

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

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

Our data consists of n i.i.d. datapoints $(Y_i, X^{(i)})$ of the variables Y and X, modelled with the classical regression equation

High-Dimensional Datasets are vast

Our data consists of n i.i.d. datapoints $(Y_i, X^{(i)})$ of the variables Y and X, modelled with the classical regression equation

$$Y = f(X^{(i)}) + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

High-Dimensional Datasets are vast

Our data consists of n i.i.d. datapoints $(Y_i, X^{(i)})$ of the variables Y and X, modelled with the classical regression equation

$$Y = f(X^{(i)}) + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

where $f : [0,1]^p \to \mathbb{R}$ is some (unknown) function and $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are independent random variables with zero expectation.

High-Dimensional Datasets are vast

Our data consists of n i.i.d. datapoints $(Y_i, X^{(i)})$ of the variables Y and X, modelled with the classical regression equation

$$Y = f(X^{(i)}) + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

where $f : [0,1]^p \to \mathbb{R}$ is some (unknown) function and $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are independent random variables with zero expectation.

Assuming that f is *smooth*, it is natural to estimate

$$f(x)$$

by some average of the Y_i associated to the $X^{(i)}$ in the vicinity (neighborhood) of x.

High-Dimensional Datasets are vast

The most simple approach is the k-*nearest neighbor estimator*,

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.

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 $||X^{(i)} - x||$

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).

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.

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

This is something you can observe in a simulation.

High-Dimensional Datasets are vast

This is something you can observe in a simulation. Consider the histograms of the sets

$$\left\{ ||X^{(i)} - X^{(j)}|| \,:\, 1 \leq i < j \leq n \right\}$$

High-Dimensional Datasets are vast

This is something you can observe in a simulation. Consider the histograms of the sets

$$\left\{ ||X^{(i)} - X^{(j)}|| \,:\, 1 \leq i < j \leq n \right\}$$

for $n = 100$ and dimensions $p = 2, 10, 100$ and 1000.

High-Dimensional Datasets are vast

This is something you can observe in a simulation. Consider the histograms of the sets

$$\left\{ ||X^{(i)} - X^{(j)}|| \,:\, 1 \leq i < j \leq n \right\}$$

for $n = 100$ and dimensions $p = 2, 10, 100$ and 1000.

When the dimension p increases, we see in the histogram that

High-Dimensional Datasets are vast

This is something you can observe in a simulation. Consider the histograms of the sets

$$\left\{ ||X^{(i)} - X^{(j)}|| \,:\, 1 \leq i < j \leq n \right\}$$

for $n = 100$ and dimensions $p = 2, 10, 100$ and 1000.

When the dimension p increases, we see in the histogram that

- the minimal distance between two points increases;

High-Dimensional Datasets are vast

This is something you can observe in a simulation. Consider the histograms of the sets

$$\left\{ ||X^{(i)} - X^{(j)}|| \,:\; 1 \leq i < j \leq n \right\}$$

for $n = 100$ and dimensions $p = 2, 10, 100$ and 1000.

When the dimension p increases, we see in the histogram that

- the minimal distance between two points increases;
- all the points are at a similar distance from one-another,

High-Dimensional Datasets are vast

This is something you can observe in a simulation. Consider the histograms of the sets

$$\left\{||X^{(i)} - X^{(j)}|| \,:\, 1 \leq i < j \leq n\right\}$$

for $n = 100$ and dimensions $p = 2, 10, 100$ and 1000.

When the dimension p increases, we see in the histogram that

- the minimal distance between two points increases;
- all the points are at a similar distance from one-another, so the notion of "nearest-points" disappears.

High-Dimensional Datasets are vast

This is something you can observe in a simulation. Consider the histograms of the sets

$$\left\{||X^{(i)} - X^{(j)}|| \,:\, 1 \leq i < j \leq n\right\}$$

for $n = 100$ and dimensions $p = 2, 10, 100$ and 1000.

When the dimension p increases, we see in the histogram that

- the minimal distance between two points increases;
- all the points are at a similar distance from one-another, so the notion of "nearest-points" disappears.

In particular, any estimator based on local averaging will fail.

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:

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:

$$\mathrm{E}\left(||X^{(i)} - X^{(j)}||^2\right) \quad =$$

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:

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

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:

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

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:

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

High-Dimensional Datasets are vast

Since $X_k^{(i)}$ and $X_k^{(j)}$ are i.i.d. uniformly distributed random variables on $[0, 1]$, we see that:

High-Dimensional Datasets are vast

Since $X_k^{(i)}$ and $X_k^{(j)}$ are i.i.d. uniformly distributed random variables on $[0,1]$, we see that:

$$\mathrm{E}\left(||X^{(i)} - X^{(j)}||^2\right) =$$

High-Dimensional Datasets are vast

Since $X_k^{(i)}$ and $X_k^{(j)}$ are i.i.d. uniformly distributed random variables on $[0,1]$, we see that:

$$\mathrm{E}\left(||X^{(i)} - X^{(j)}||^2\right) = p \cdot \mathrm{E}\left((U - U')^2\right) =$$

High-Dimensional Datasets are vast

Since $X_k^{(i)}$ and $X_k^{(j)}$ are i.i.d. uniformly distributed random variables on $[0,1]$, we see that:

$$\mathrm{E}\left(||X^{(i)} - X^{(j)}||^2\right) = p \cdot \mathrm{E}\left((U - U')^2\right) = p/6.$$

High-Dimensional Datasets are vast

Since $X_k^{(i)}$ and $X_k^{(j)}$ are i.i.d. uniformly distributed random variables on $[0,1]$, we see that:

$$\mathrm{E}\left(||X^{(i)} - X^{(j)}||^2\right) = p \cdot \mathrm{E}\left((U - U')^2\right) = p/6.$$

The last two steps are an *exercise* for you.

High-Dimensional Datasets are vast

Since $X_k^{(i)}$ and $X_k^{(j)}$ are i.i.d. uniformly distributed random variables on $[0,1]$, we see that:

$$\mathrm{E}\left(||X^{(i)} - X^{(j)}||^2\right) = p \cdot \mathrm{E}\left((U - U')^2\right) = p/6.$$

The last two steps are an *exercise* for you.

The variance of this mean square distance is:

High-Dimensional Datasets are vast

Since $X_k^{(i)}$ and $X_k^{(j)}$ are i.i.d. uniformly distributed random variables on $[0, 1]$, we see that:

$$\mathrm{E}\left(||X^{(i)} - X^{(j)}||^2\right) = p \cdot \mathrm{E}\left((U - U')^2\right) = p/6.$$

The last two steps are an *exercise* for you.

The variance of this mean square distance is:

$$\mathrm{Var}\left(||X^{(i)} - X^{(j)}||^2\right) \quad =$$

High-Dimensional Datasets are vast

Since $X_k^{(i)}$ and $X_k^{(j)}$ are i.i.d. uniformly distributed random variables on $[0,1]$, we see that:

$$\mathrm{E}\left(||X^{(i)} - X^{(j)}||^2\right) = p \cdot \mathrm{E}\left((U - U')^2\right) = p/6.$$

The last two steps are an *exercise* for you.

The variance of this mean square distance is:

$$\begin{aligned}\mathrm{Var}\left(||X^{(i)} - X^{(j)}||^2\right) &= \mathrm{Var}\left((X_1^{(i)} - X_1^{(j)})^2 + \cdots + (X_p^{(i)} - X_p^{(j)})^2\right) \\ &= \end{aligned}$$

High-Dimensional Datasets are vast

Since $X_k^{(i)}$ and $X_k^{(j)}$ are i.i.d. uniformly distributed random variables on $[0,1]$, we see that:

$$\mathrm{E}\left(||X^{(i)} - X^{(j)}||^2\right) = p \cdot \mathrm{E}\left((U - U')^2\right) = p/6.$$

The last two steps are an *exercise* for you.

The variance of this mean square distance is:

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

High-Dimensional Datasets are vast

Since $X_k^{(i)}$ and $X_k^{(j)}$ are i.i.d. uniformly distributed random variables on $[0,1]$, we see that:

$$\mathrm{E}\left(||X^{(i)} - X^{(j)}||^2\right) = p \cdot \mathrm{E}\left((U - U')^2\right) = p/6.$$

The last two steps are an *exercise* for you.

The variance of this mean square distance is:

$$\begin{aligned}
\mathrm{Var}\left(||X^{(i)} - X^{(j)}||^2\right) &= \mathrm{Var}\left((X_1^{(i)} - X_1^{(j)})^2 + \cdots + (X_p^{(i)} - X_p^{(j)})^2\right) \\
&= \sum_{k=1}^{p} \mathrm{Var}\left((X_k^{(i)} - X_k^{(j)})^2\right) \\
&= p \cdot \mathrm{Var}\left((U - U')^2\right) \\
&\approx
\end{aligned}$$

High-Dimensional Datasets are vast

Since $X_k^{(i)}$ and $X_k^{(j)}$ are i.i.d. uniformly distributed random variables on $[0,1]$, we see that:

$$\mathrm{E}\left(||X^{(i)} - X^{(j)}||^2\right) = p \cdot \mathrm{E}\left((U - U')^2\right) = p/6.$$

The last two steps are an *exercise* for you.

The variance of this mean square distance is:

$$\begin{aligned}
\mathrm{Var}\left(||X^{(i)} - X^{(j)}||^2\right) &= \mathrm{Var}\left((X_1^{(i)} - X_1^{(j)})^2 + \cdots + (X_p^{(i)} - X_p^{(j)})^2\right) \\
&= \sum_{k=1}^{p} \mathrm{Var}\left((X_k^{(i)} - X_k^{(j)})^2\right) \\
&= p \cdot \mathrm{Var}\left((U - U')^2\right) \\
&\approx 0.04p
\end{aligned}$$

High-Dimensional Datasets are vast

Since $X_k^{(i)}$ and $X_k^{(j)}$ are i.i.d. uniformly distributed random variables on $[0,1]$, we see that:

$$\mathrm{E}\left(||X^{(i)} - X^{(j)}||^2\right) = p \cdot \mathrm{E}\left((U - U')^2\right) = p/6.$$

The last two steps are an *exercise* for you.

The variance of this mean square distance is:

$$\begin{aligned} \mathrm{Var}\left(||X^{(i)} - X^{(j)}||^2\right) &= \mathrm{Var}\left((X_1^{(i)} - X_1^{(j)})^2 + \cdots + (X_p^{(i)} - X_p^{(j)})^2\right) \\ &= \sum_{k=1}^{p} \mathrm{Var}\left((X_k^{(i)} - X_k^{(j)})^2\right) \\ &= p \cdot \mathrm{Var}\left((U - U')^2\right) \\ &\approx 0.04p \end{aligned}$$

(Note that the second step is due to independence.

High-Dimensional Datasets are vast

Since $X_k^{(i)}$ and $X_k^{(j)}$ are i.i.d. uniformly distributed random variables on $[0,1]$, we see that:

$$\mathrm{E}\left(||X^{(i)} - X^{(j)}||^2\right) = p \cdot \mathrm{E}\left((U - U')^2\right) = p/6.$$

The last two steps are an *exercise* for you.

The variance of this mean square distance is:

$$\begin{aligned}
\mathrm{Var}\left(||X^{(i)} - X^{(j)}||^2\right) &= \mathrm{Var}\left((X_1^{(i)} - X_1^{(j)})^2 + \cdots + (X_p^{(i)} - X_p^{(j)})^2\right) \\
&= \sum_{k=1}^{p} \mathrm{Var}\left((X_k^{(i)} - X_k^{(j)})^2\right) \\
&= p \cdot \mathrm{Var}\left((U - U')^2\right) \\
&\approx 0.04p
\end{aligned}$$

(Note that the second step is due to independence. Again the last step is an *exercise* for you).

High-Dimensional Datasets are vast

So the standard deviation of this mean square distance is:

High-Dimensional Datasets are vast

So the standard deviation of this mean square distance is:

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

High-Dimensional Datasets are vast

So the standard deviation of this mean square distance is:

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

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

High-Dimensional Datasets are vast

So the standard deviation of this mean square distance is:

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

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

High-Dimensional Datasets are vast

So the standard deviation of this mean square distance is:

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

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

$$\frac{\mathrm{sdev}\left(||X^{(i)} - X^{(j)}||^2\right)}{\mathrm{E}\left(||X^{(i)} - X^{(j)}||^2\right)} \approx$$

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

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

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

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

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

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

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

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?

A simple remedy for the observations we just made (that the observations become isolated in the sample space due to dimensionality) seems obvious:

How many observations do we need?

A simple remedy for the observations we just made (that the observations become isolated in the sample space due to dimensionality) seems obvious: *just increase the number of observations* n *and the isolation will disappear*.

How many observations do we need?

A simple remedy for the observations we just made (that the observations become isolated in the sample space due to dimensionality) seems obvious: *just increase the number of observations* n *and the isolation will disappear*.

The sheer vastness of the sample space might already be an indication that this thought is too simple, but let's try to quantify it a little bit more!

How many observations do we need?

A simple remedy for the observations we just made (that the observations become isolated in the sample space due to dimensionality) seems obvious: *just increase the number of observations* n *and the isolation will disappear*.

The sheer vastness of the sample space might already be an indication that this thought is too simple, but let's try to quantify it a little bit more!

Suppose that for *every* $x \in [0,1]^p$ we have at least one $X^{(i)}$ (from the observations $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$) which is at distance 1 or less from x.

How many observations do we need?

A simple remedy for the observations we just made (that the observations become isolated in the sample space due to dimensionality) seems obvious: *just increase the number of observations* n *and the isolation will disappear*.

The sheer vastness of the sample space might already be an indication that this thought is too simple, but let's try to quantify it a little bit more!

Suppose that for *every* $x \in [0,1]^p$ we have at least one $X^{(i)}$ (from the observations $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$) which is at distance 1 or less from x. How large should n then be *at least*?

How many observations do we need?

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

How many observations do we need?

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

$$V_p(r) =$$

How many observations do we need?

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

$$V_p(r) = \frac{\pi^{p/2}}{\Gamma(p/2+1)} \approx$$

How many observations do we need?

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

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

How many observations do we need?

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

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

How many observations do we need?

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

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

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

How many observations do we need?

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

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

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?

If $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ are such that for any $x \in [0,1]^p$ there exists at least one $x^{(i)}$ such that

$$||x^{(i)} - x|| \leq 1,$$

How many observations do we need?

If $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ are such that for any $x \in [0,1]^p$ there exists at least one $x^{(i)}$ such that

$$||x^{(i)} - x|| \leq 1,$$

then the hypercube $[0,1]^p$ is contained in the union of the (closed) hyperballs centered in $x^{(1)}, x^{(2)}, \dots, x^{(n)}$. I.e.

How many observations do we need?

If $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ are such that for any $x \in [0,1]^p$ there exists at least one $x^{(i)}$ such that

$$||x^{(i)} - x|| \leq 1,$$

then the hypercube $[0,1]^p$ is contained in the union of the (closed) hyperballs centered in $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$. I.e.

$$[0,1]^p \subset \bigcup_{i=1}^{n} B_p(x^{(i)}, 1).$$

How many observations do we need?

As a consequence,

How many observations do we need?

As a consequence,

$$1 = V_p\left([0,1]^p\right) \leq$$

How many observations do we need?

As a consequence,

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

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:

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:

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

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.

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.

$$n \geq \frac{\Gamma(p/2+1)}{\pi^{p/2}} \approx$$

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.

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

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.

How many observations do we need?

This would render an amount of observations totally unrealistic, as the following table shows:

How many observations do we need?

This would render an amount of observations totally unrealistic, as the following table shows:

p	20	30	50	100	150	200
n						

How many observations do we need?

This would render an amount of observations totally unrealistic, as the following table shows:

p	20	30	50	100	150	200
n	39					

How many observations do we need?

This would render an amount of observations totally unrealistic, as the following table shows:

p	20	30	50	100	150	200
n	39	45630				

How many observations do we need?

This would render an amount of observations totally unrealistic, as the following table shows:

p	20	30	50	100	150	200
n	39	45630	$5.7 \cdot 10^{12}$			

How many observations do we need?

This would render an amount of observations totally unrealistic, as the following table shows:

p	20	30	50	100	150	200
n	39	45630	$5.7 \cdot 10^{12}$	$42 \cdot 10^{39}$		

How many observations do we need?

This would render an amount of observations totally unrealistic, as the following table shows:

p	20	30	50	100	150	200
n	39	45630	$5.7 \cdot 10^{12}$	$42 \cdot 10^{39}$	$1.28 \cdot 10^{72}$	

How many observations do we need?

This would render an amount of observations totally unrealistic, as the following table shows:

p	20	30	50	100	150	200
n	39	45630	$5.7 \cdot 10^{12}$	$42 \cdot 10^{39}$	$1.28 \cdot 10^{72}$	larger than the estimated number of particles in the observable universe

How many observations do we need?

This would render an amount of observations totally unrealistic, as the following table shows:

p	20	30	50	100	150	200
n	39	45630	$5.7 \cdot 10^{12}$	$42 \cdot 10^{39}$	$1.28 \cdot 10^{72}$	larger than the estimated number of particles in the observable universe :)

How many observations do we need?

This would render an amount of observations totally unrealistic, as the following table shows:

p	20	30	50	100	150	200
n	39	45630	$5.7 \cdot 10^{12}$	$42 \cdot 10^{39}$	$1.28 \cdot 10^{72}$	larger than the estimated number of particles in the observable universe :)

Mind you: these values for n are only **lower bounds** of n;

How many observations do we need?

This would render an amount of observations totally unrealistic, as the following table shows:

p	20	30	50	100	150	200
n	39	45630	$5.7 \cdot 10^{12}$	$42 \cdot 10^{39}$	$1.28 \cdot 10^{72}$	larger than the estimated number of particles in the observable universe :)

Mind you: these values for n are only **lower bounds** of n; in reality one would probably need more balls to "cover" the hypercube $[0,1]^p$, simply because the balls are not so "nicely spread-out" over $[0,1]^p$.

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.

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

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 $\mathrm{Var}(\varepsilon_1) = \sigma^2 (> 0)$.

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:

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:

$$|F(X_1) - F(\theta_1)| \quad =$$

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:

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

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:

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

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:

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

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

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 $\mathbb{R}^+$):

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 $\mathbb{R}^+$):

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

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:

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:

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

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:

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

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 assume we need to estimate the p-dimensional function $F(\theta_1, \theta_2, \ldots, \theta_p)$ from the noisy observations $X_j = \theta_j + \varepsilon_j$ of the θ_j.

Fluctuation accumulate

Now assume we need to estimate the p-dimensional function $F(\theta_1, \theta_2, \ldots, \theta_p)$ from the noisy observations $X_j = \theta_j + \varepsilon_j$ of the θ_j.

Assume that the noise variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ are independent and all have expectation zero and variance σ^2.

Fluctuation accumulate

Now assume we need to estimate the p-dimensional function $F(\theta_1, \theta_2, \ldots, \theta_p)$ from the noisy observations $X_j = \theta_j + \varepsilon_j$ of the θ_j.

Assume that the noise variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ are independent and all have expectation zero and variance σ^2.

If (as in the one-dimensional case) we have that F is 1-Lipschitz, one finds:

Fluctuation accumulate

Now assume we need to estimate the p-dimensional function $F(\theta_1, \theta_2, \ldots, \theta_p)$ from the noisy observations $X_j = \theta_j + \varepsilon_j$ of the θ_j.

Assume that the noise variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ are independent and all have expectation zero and variance σ^2.

If (as in the one-dimensional case) we have that F is 1-Lipschitz, one finds:

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

Fluctuation accumulate

Now assume we need to estimate the p-dimensional function $F(\theta_1, \theta_2, \ldots, \theta_p)$ from the noisy observations $X_j = \theta_j + \varepsilon_j$ of the θ_j.

Assume that the noise variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ are independent and all have expectation zero and variance σ^2.

If (as in the one-dimensional case) we have that F is 1-Lipschitz, one finds:

$$\begin{aligned} \mathrm{E}\left(||F(X_1, \ldots, X_p) - F(\theta_1, \theta_2, \ldots, \theta_p)||^2\right) &\leq \mathrm{E}\left(||(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p)||^2\right) \\ &= \end{aligned}$$

Fluctuation accumulate

Now assume we need to estimate the p-dimensional function $F(\theta_1, \theta_2, \ldots, \theta_p)$ from the noisy observations $X_j = \theta_j + \varepsilon_j$ of the θ_j.

Assume that the noise variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ are independent and all have expectation zero and variance σ^2.

If (as in the one-dimensional case) we have that F is 1-Lipschitz, one finds:

$$\begin{aligned} \mathrm{E}\left(||F(X_1, \ldots, X_p) - F(\theta_1, \theta_2, \ldots, \theta_p)||^2\right) &\leq \mathrm{E}\left(||(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p)||^2\right) \\ &= \mathrm{E}\left(\varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_p^2\right) \\ \\ &= \end{aligned}$$

Fluctuation accumulate

Now assume we need to estimate the p-dimensional function $F(\theta_1, \theta_2, \ldots, \theta_p)$ from the noisy observations $X_j = \theta_j + \varepsilon_j$ of the θ_j.

Assume that the noise variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ are independent and all have expectation zero and variance σ^2.

If (as in the one-dimensional case) we have that F is 1-Lipschitz, one finds:

$$\begin{aligned}
\mathrm{E}\left(||F(X_1,\ldots,X_p) - F(\theta_1,\theta_2,\ldots,\theta_p)||^2\right) &\leq \mathrm{E}\left(||(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_p)||^2\right) \\
&= \mathrm{E}\left(\varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_p^2\right) \\
&= \sum_{j=1}^{p} \mathrm{E}\left(\varepsilon_j^2\right) \\
&=
\end{aligned}$$

Fluctuation accumulate

Now assume we need to estimate the p-dimensional function $F(\theta_1, \theta_2, \dots, \theta_p)$ from the noisy observations $X_j = \theta_j + \varepsilon_j$ of the θ_j.

Assume that the noise variables $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$ are independent and all have expectation zero and variance σ^2.

If (as in the one-dimensional case) we have that F is 1-Lipschitz, one finds:

$$\begin{aligned}
\mathrm{E}\left(||F(X_1, \dots, X_p) - F(\theta_1, \theta_2, \dots, \theta_p)||^2\right) &\leq \mathrm{E}\left(||(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p)||^2\right) \\
&= \mathrm{E}\left(\varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_p^2\right) \\
&= \sum_{j=1}^{p} \mathrm{E}\left(\varepsilon_j^2\right) \\
&= \sum_{j=1}^{p} \sigma^2 \\
&=
\end{aligned}$$

Fluctuation accumulate

Now assume we need to estimate the p-dimensional function $F(\theta_1, \theta_2, \ldots, \theta_p)$ from the noisy observations $X_j = \theta_j + \varepsilon_j$ of the θ_j.

Assume that the noise variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ are independent and all have expectation zero and variance σ^2.

If (as in the one-dimensional case) we have that F is 1-Lipschitz, one finds:

$$\begin{aligned}
\mathrm{E}\left(||F(X_1, \ldots, X_p) - F(\theta_1, \theta_2, \ldots, \theta_p)||^2\right) &\leq \mathrm{E}\left(||(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p)||^2\right) \\
&= \mathrm{E}\left(\varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_p^2\right) \\
&= \sum_{j=1}^{p} \mathrm{E}\left(\varepsilon_j^2\right) \\
&= \sum_{j=1}^{p} \sigma^2 \\
&= p \cdot \sigma^2.
\end{aligned}$$

Fluctuation accumulate

Now assume we need to estimate the p-dimensional function $F(\theta_1, \theta_2, \dots, \theta_p)$ from the noisy observations $X_j = \theta_j + \varepsilon_j$ of the θ_j.

Assume that the noise variables $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$ are independent and all have expectation zero and variance σ^2.

If (as in the one-dimensional case) we have that F is 1-Lipschitz, one finds:

$$\begin{aligned}
\mathrm{E}\left(||F(X_1, \dots, X_p) - F(\theta_1, \theta_2, \dots, \theta_p)||^2\right) &\leq \mathrm{E}\left(||(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p)||^2\right) \\
&= \mathrm{E}\left(\varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_p^2\right) \\
&= \sum_{j=1}^{p} \mathrm{E}\left(\varepsilon_j^2\right) \\
&= \sum_{j=1}^{p} \sigma^2 \\
&= p \cdot \sigma^2.
\end{aligned}$$

So if p becomes large, $p\sigma^2$ is getting pretty big too!

Fluctuation accumulate

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

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

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!

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:

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:

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

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:

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

but then the mean square error error

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:

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

but then the mean square error error

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

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:

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

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 $p\sigma^2$.

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:

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

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

Assume we have n observations

High-Dimensional Linear Regression

Assume we have n observations

$$Y_i = \left\langle x^{(i)}, \beta^\star \right\rangle + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

High-Dimensional Linear Regression

Assume we have n observations

$$Y_i = \left\langle x^{(i)}, \beta^\star \right\rangle + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

with the responce $Y_i \in \mathbb{R}$ and the covariates $x^{(i)} \in \mathbb{R}^p$.

High-Dimensional Linear Regression

Assume we have n observations

$$Y_i = \left\langle x^{(i)}, \beta^\star \right\rangle + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

with the responce $Y_i \in \mathbb{R}$ and the covariates $x^{(i)} \in \mathbb{R}^p$.

We want to estimate (the "true") $\beta^\star \in \mathbb{R}^p$, and we assume that $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random variables with zero mean and variance $\sigma^2 > 0$.

High-Dimensional Linear Regression

Assume we have n observations

$$Y_i = \left\langle x^{(i)}, \beta^\star \right\rangle + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

with the responce $Y_i \in \mathbb{R}$ and the covariates $x^{(i)} \in \mathbb{R}^p$.

We want to estimate (the "true") $\beta^\star \in \mathbb{R}^p$, and we assume that $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random variables with zero mean and variance $\sigma^2 > 0$.

Writing

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},$$

High-Dimensional Linear Regression

Assume we have n observations

$$Y_i = \left\langle x^{(i)}, \beta^\star \right\rangle + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

with the responce $Y_i \in \mathbb{R}$ and the covariates $x^{(i)} \in \mathbb{R}^p$.

We want to estimate (the "true") $\beta^\star \in \mathbb{R}^p$, and we assume that $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random variables with zero mean and variance $\sigma^2 > 0$.

Writing

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(n)})^T \end{pmatrix} \quad \text{and}$$

High-Dimensional Linear Regression

Assume we have n observations

$$Y_i = \left\langle x^{(i)}, \beta^\star \right\rangle + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

with the responce $Y_i \in \mathbb{R}$ and the covariates $x^{(i)} \in \mathbb{R}^p$.

We want to estimate (the "true") $\beta^\star \in \mathbb{R}^p$, and we assume that $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random variables with zero mean and variance $\sigma^2 > 0$.

Writing

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(n)})^T \end{pmatrix} \text{ and } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

High-Dimensional Linear Regression

Assume we have n observations

$$Y_i = \left\langle x^{(i)}, \beta^\star \right\rangle + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

with the responce $Y_i \in \mathbb{R}$ and the covariates $x^{(i)} \in \mathbb{R}^p$.

We want to estimate (the "true") $\beta^\star \in \mathbb{R}^p$, and we assume that $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random variables with zero mean and variance $\sigma^2 > 0$.

Writing

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(n)})^T \end{pmatrix} \text{ and } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

we have

High-Dimensional Linear Regression

Assume we have n observations

$$Y_i = \left\langle x^{(i)}, \beta^\star \right\rangle + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

with the responce $Y_i \in \mathbb{R}$ and the covariates $x^{(i)} \in \mathbb{R}^p$.

We want to estimate (the "true") $\beta^\star \in \mathbb{R}^p$, and we assume that $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random variables with zero mean and variance $\sigma^2 > 0$.

Writing

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(n)})^T \end{pmatrix} \text{ and } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

we have $Y = \mathbf{X}\beta^\star + \varepsilon$.

High-Dimensional Linear Regression

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

High-Dimensional Linear Regression

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

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

High-Dimensional Linear Regression

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

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

which is uniquely defined if the rank of $\mathbf{X}$ is p.

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.

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:

High-Dimensional Linear Regression

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

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

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:

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

High-Dimensional Linear Regression

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

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

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:

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:

$$\mathrm{E}\left(||\hat{\beta} - \beta^\star||^2\right) =$$

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:

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

High-Dimensional Linear Regression

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

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

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:

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

High-Dimensional Linear Regression

So if the columns of $\mathbf{X}$ are orthonormal

High-Dimensional Linear Regression

So if the columns of $\mathbf{X}$ are orthonormal (i.e. 2-by-2 perpendicular and each of length 1),

High-Dimensional Linear Regression

So if the columns of $\mathbf{X}$ are orthonormal (i.e. 2-by-2 perpendicular and each of length 1), then

High-Dimensional Linear Regression

So if the columns of $\mathbf{X}$ are orthonormal (i.e. 2-by-2 perpendicular and each of length 1), then

$$\mathbf{X}^T\mathbf{X} = \mathrm{Id},$$

High-Dimensional Linear Regression

So if the columns of $\mathbf{X}$ are orthonormal (i.e. 2-by-2 perpendicular and each of length 1), then

$$\mathbf{X}^T\mathbf{X} = \mathrm{Id},$$

and we immediately find that

High-Dimensional Linear Regression

So if the columns of $\mathbf{X}$ are orthonormal (i.e. 2-by-2 perpendicular and each of length 1), then

$$\mathbf{X}^T\mathbf{X} = \mathrm{Id},$$

and we immediately find that

$$\mathrm{E}\left(||\hat{\beta} - \beta^\star||^2\right) =$$

High-Dimensional Linear Regression

So if the columns of $\mathbf{X}$ are orthonormal (i.e. 2-by-2 perpendicular and each of length 1), then

$$\mathbf{X}^T\mathbf{X} = \mathrm{Id},$$

and we immediately find that

$$\mathrm{E}\left(||\hat{\beta} - \beta^{\star}||^2\right) = p \cdot \sigma^2.$$

High-Dimensional Linear Regression

So if the columns of $\mathbf{X}$ are orthonormal (i.e. 2-by-2 perpendicular and each of length 1), then

$$\mathbf{X}^T\mathbf{X} = \mathrm{Id},$$

and we immediately find that

$$\mathrm{E}\left(||\hat{\beta} - \beta^\star||^2\right) = p \cdot \sigma^2.$$

But then the estimation error grows linearly with the dimension p of the covariate $x^{(i)}$!

Computational Complexity

Another burden arises in high-dimensional settings:

Computational Complexity

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

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

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

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.

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.

Computational Complexity

So we might be inclined to compare the outcomes of the family of regression problems

$$y = \sum_{j \in m} \beta_j^\star x_j + \varepsilon \qquad \text{for each } m \subset \{1, 2, \ldots, p\}.$$

Computational Complexity

So we might be inclined to compare the outcomes of the family of regression problems

$$y = \sum_{j \in m} \beta_j^\star x_j + \varepsilon \qquad \text{for each } m \subset \{1, 2, \ldots, p\}. \tag{3}$$

Unfortunately, the cardinality of $\{m : \; m \subset \{1, 2, \ldots, p\}\}$ is 2^p;

Computational Complexity

So we might be inclined to compare the outcomes of the family of regression problems

$$y = \sum_{j \in m} \beta_j^\star x_j + \varepsilon \qquad \text{for each } m \subset \{1, 2, \ldots, p\}. \tag{3}$$

Unfortunately, the cardinality of $\{m : \; m \subset \{1, 2, \ldots, p\}\}$ is 2^p; it grows exponentially with p!!

Computational Complexity

So we might be inclined to compare the outcomes of the family of regression problems

$$y = \sum_{j \in m} \beta_j^\star x_j + \varepsilon \qquad \text{for each } m \subset \{1, 2, \dots, p\}. \tag{3}$$

Unfortunately, the cardinality of $\{m : \ m \subset \{1, 2, \dots, p\}\}$ is 2^p; it grows exponentially with p!!

So when p is 10 or more, it becomes hard to impossible to calculate the 2^p estimators $\hat{\beta_m}$ associated to the model (3).

Computational Complexity

So we might be inclined to compare the outcomes of the family of regression problems

$$y = \sum_{j \in m} \beta_j^\star x_j + \varepsilon \qquad \text{for each } m \subset \{1, 2, \ldots, p\}. \tag{3}$$

Unfortunately, the cardinality of $\{m : \ m \subset \{1, 2, \ldots, p\}\}$ is 2^p; it grows exponentially with p!!

So when p is 10 or more, it becomes hard to impossible to calculate the 2^p estimators $\hat{\beta_m}$ associated to the model (3).

For example, $2^{10} =$

Computational Complexity

So we might be inclined to compare the outcomes of the family of regression problems

$$y = \sum_{j \in m} \beta_j^{\star} x_j + \varepsilon \qquad \text{for each } m \subset \{1, 2, \ldots, p\}. \tag{3}$$

Unfortunately, the cardinality of $\{m : \ m \subset \{1, 2, \ldots, p\}\}$ is 2^p; it grows exponentially with p!!

So when p is 10 or more, it becomes hard to impossible to calculate the 2^p estimators $\hat{\beta_m}$ associated to the model (3).

For example, $2^{10} = 1024$,

Computational Complexity

So we might be inclined to compare the outcomes of the family of regression problems

$$y = \sum_{j \in m} \beta_j^\star x_j + \varepsilon \qquad \text{for each } m \subset \{1, 2, \ldots, p\}. \tag{3}$$

Unfortunately, the cardinality of $\{m : \; m \subset \{1, 2, \ldots, p\}\}$ is 2^p; it grows exponentially with p!!

So when p is 10 or more, it becomes hard to impossible to calculate the 2^p estimators $\hat{\beta_m}$ associated to the model (3).

For example, $2^{10} = 1024$, $2^{20} =$

Computational Complexity

So we might be inclined to compare the outcomes of the family of regression problems

$$y = \sum_{j \in m} \beta_j^\star x_j + \varepsilon \qquad \text{for each } m \subset \{1, 2, \ldots, p\}. \tag{3}$$

Unfortunately, the cardinality of $\{m : \ m \subset \{1, 2, \ldots, p\}\}$ is 2^p; it grows exponentially with p!!

So when p is 10 or more, it becomes hard to impossible to calculate the 2^p estimators $\hat{\beta_m}$ associated to the model (3).

For example, $2^{10} = 1024$, $2^{20} = 1\,048\,576$, while

Computational Complexity

So we might be inclined to compare the outcomes of the family of regression problems

$$y = \sum_{j \in m} \beta_j^\star x_j + \varepsilon \qquad \text{for each } m \subset \{1, 2, \ldots, p\}. \tag{3}$$

Unfortunately, the cardinality of $\{m : \; m \subset \{1, 2, \ldots, p\}\}$ is 2^p; it grows exponentially with p!!

So when p is 10 or more, it becomes hard to impossible to calculate the 2^p estimators $\hat{\beta_m}$ associated to the model (3).

For example, $2^{10} = 1024$, $2^{20} = 1\,048\,576$, while $2^{30} =$

Computational Complexity

So we might be inclined to compare the outcomes of the family of regression problems

$$y = \sum_{j \in m} \beta_j^{\star} x_j + \varepsilon \qquad \text{for each } m \subset \{1, 2, \ldots, p\}. \tag{3}$$

Unfortunately, the cardinality of $\{m : \; m \subset \{1, 2, \ldots, p\}\}$ is 2^p; it grows exponentially with p!!

So when p is 10 or more, it becomes hard to impossible to calculate the 2^p estimators $\hat{\beta_m}$ associated to the model (3).

For example, $2^{10} = 1024$, $2^{20} = 1\,048\,576$, while $2^{30} = 1\,073\,741\,824\ \ldots$

Computational Complexity

So we might be inclined to compare the outcomes of the family of regression problems

$$y = \sum_{j \in m} \beta_j^\star x_j + \varepsilon \qquad \text{for each } m \subset \{1, 2, \ldots, p\}. \tag{3}$$

Unfortunately, the cardinality of $\{m: \ m \subset \{1, 2, \ldots, p\}\}$ is 2^p; it grows exponentially with p!!

So when p is 10 or more, it becomes hard to impossible to calculate the 2^p estimators $\hat{\beta_m}$ associated to the model (3).

For example, $2^{10} = 1024$, $2^{20} = 1\,048\,576$, while $2^{30} = 1\,073\,741\,824$ ... a burden indeed!

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

Gaussian distributions are known to have very thin tails.

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

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

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

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

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

of a standard Gaussian distribution (i.e. a $N(0, I_p)$-distributed random variable)

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

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

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.

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

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

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 . . .

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

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

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!!

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

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

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

First note that

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

First note that

$$g_p(0) = \frac{1}{(\sqrt{2\pi})^p},$$

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

First note that

$$g_p(0) = \frac{1}{(\sqrt{2\pi})^p},$$

which decays exponentially fast to 0 as $p \to \infty$.

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

First note that

$$g_p(0) = \frac{1}{(\sqrt{2\pi})^p},$$

which decays exponentially fast to 0 as $p \to \infty$. So the Gaussian distribution in high-dimensions is quite "flat,"

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

First note that

$$g_p(0) = \frac{1}{(\sqrt{2\pi})^p},$$

which decays exponentially fast to 0 as $p \to \infty$. So the Gaussian distribution in high-dimensions is quite "flat," much more than in dimension $p = 1$ or $p = 2$.

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

First note that

$$g_p(0) = \frac{1}{(\sqrt{2\pi})^p},$$

which decays exponentially fast to 0 as $p \to \infty$. So the Gaussian distribution in high-dimensions is quite "flat," much more than in dimension $p = 1$ or $p = 2$.

Just to get an idea where the mass is located, let's compute the mass in the "bell"

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

First note that

$$g_p(0) = \frac{1}{(\sqrt{2\pi})^p},$$

which decays exponentially fast to 0 as $p \to \infty$. So the Gaussian distribution in high-dimensions is quite "flat," much more than in dimension $p = 1$ or $p = 2$.

Just to get an idea where the mass is located, let's compute the mass in the "bell" (the central part where the density is the largest).

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

First note that

$$g_p(0) = \frac{1}{(\sqrt{2\pi})^p},$$

which decays exponentially fast to 0 as $p \to \infty$. So the Gaussian distribution in high-dimensions is quite "flat," much more than in dimension $p = 1$ or $p = 2$.

Just to get an idea where the mass is located, let's compute the mass in the "bell" (the central part where the density is the largest). Let $\delta > 0$ be a small positive number and write

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

First note that

$$g_p(0) = \frac{1}{(\sqrt{2\pi})^p},$$

which decays exponentially fast to 0 as $p \to \infty$. So the Gaussian distribution in high-dimensions is quite "flat," much more than in dimension $p = 1$ or $p = 2$.

Just to get an idea where the mass is located, let's compute the mass in the "bell" (the central part where the density is the largest). Let $\delta > 0$ be a small positive number and write

$$B_{p,\delta} = \{x \in \mathbb{R}^p \; : \; g_p(x) \geq \delta g_p(0)\}.$$

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

First note that

$$g_p(0) = \frac{1}{(\sqrt{2\pi})^p},$$

which decays exponentially fast to 0 as $p \to \infty$. So the Gaussian distribution in high-dimensions is quite "flat," much more than in dimension $p = 1$ or $p = 2$.

Just to get an idea where the mass is located, let's compute the mass in the "bell" (the central part where the density is the largest). Let $\delta > 0$ be a small positive number and write

$$B_{p,\delta} = \{x \in \mathbb{R}^p \;:\; g_p(x) \geq \delta g_p(0)\}.$$

We see that

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

First note that

$$g_p(0) = \frac{1}{(\sqrt{2\pi})^p},$$

which decays exponentially fast to 0 as $p \to \infty$. So the Gaussian distribution in high-dimensions is quite "flat," much more than in dimension $p = 1$ or $p = 2$.

Just to get an idea where the mass is located, let's compute the mass in the "bell" (the central part where the density is the largest). Let $\delta > 0$ be a small positive number and write

$$B_{p,\delta} = \{x \in \mathbb{R}^p \ : \ g_p(x) \geq \delta g_p(0)\}.$$

We see that

$$B_{p,\delta} = \left\{x \in \mathbb{R}^p \ : \ ||x||^2 \leq 2\log(1/\delta)\right\}.$$

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

Recall the Markov Inequality:

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

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

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

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

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

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

Proof of the Markov Inequality

Proof of the Markov Inequality

Let $1_{\{X \geq t\}}$ be the indicator function of the set/event $\{X \geq t\}$.

Proof of the Markov Inequality

Let $1_{\{X \geq t\}}$ be the indicator function of the set/event $\{X \geq t\}$. I.e.

Proof of the Markov Inequality

Let $1_{\{X \geq t\}}$ be the indicator function of the set/event $\{X \geq t\}$. I.e.

$$1_{\{X \geq t\}}(x) = \begin{cases} 1, & \text{if } x \in \{X \geq t\} \\ 0, & \text{otherwise.} \end{cases}$$

Proof of the Markov Inequality

Let $1_{\{X\geq t\}}$ be the indicator function of the set/event $\{X \geq t\}$. I.e.

$$1_{\{X\geq t\}}(x) = \begin{cases} 1, & \text{if } x \in \{X \geq t\} \\ 0, & \text{otherwise.} \end{cases}$$

Since ψ is a positive, non-decreasing function, we have

Proof of the Markov Inequality

Let $1_{\{X \geq t\}}$ be the indicator function of the set/event $\{X \geq t\}$. I.e.

$$1_{\{X \geq t\}}(x) = \begin{cases} 1, & \text{if } x \in \{X \geq t\} \\ 0, & \text{otherwise.} \end{cases}$$

Since ψ is a positive, non-decreasing function, we have

$$P(X \geq t) \quad =$$

Proof of the Markov Inequality

Let $1_{\{X \geq t\}}$ be the indicator function of the set/event $\{X \geq t\}$. I.e.

$$1_{\{X\geq t\}}(x) = \begin{cases} 1, & \text{if } x \in \{X \geq t\} \\ 0, & \text{otherwise.} \end{cases}$$

Since ψ is a positive, non-decreasing function, we have

$$\begin{aligned} \mathrm{P}(X \geq t) &= 0 \cdot \mathrm{P}(X < t) + 1 \cdot \mathrm{P}(X \geq t) \\ &= \end{aligned}$$

Proof of the Markov Inequality

Let $1_{\{X \geq t\}}$ be the indicator function of the set/event $\{X \geq t\}$. I.e.

$$1_{\{X \geq t\}}(x) = \begin{cases} 1, & \text{if } x \in \{X \geq t\} \\ 0, & \text{otherwise.} \end{cases}$$

Since ψ is a positive, non-decreasing function, we have

$$\begin{aligned} \mathrm{P}(X \geq t) &= 0 \cdot \mathrm{P}(X < t) + 1 \cdot \mathrm{P}(X \geq t) \\ &= \mathrm{E}\left(1_{\{X \geq t\}}\right) \\ &\leq \end{aligned}$$

Proof of the Markov Inequality

Let $1_{\{X \geq t\}}$ be the indicator function of the set/event $\{X \geq t\}$. I.e.

$$1_{\{X \geq t\}}(x) = \begin{cases} 1, & \text{if } x \in \{X \geq t\} \\ 0, & \text{otherwise.} \end{cases}$$

Since ψ is a positive, non-decreasing function, we have

$$\begin{aligned} \mathrm{P}(X \geq t) &= 0 \cdot \mathrm{P}(X < t) + 1 \cdot \mathrm{P}(X \geq t) \\ &= \mathrm{E}\left(1_{\{X \geq t\}}\right) \\ &\leq \mathrm{E}\left(\frac{\psi(X)}{\psi(t)} 1_{\{X \geq t\}}\right) \\ &= \end{aligned}$$

Proof of the Markov Inequality

Let $1_{\{X \geq t\}}$ be the indicator function of the set/event $\{X \geq t\}$. I.e.

$$1_{\{X \geq t\}}(x) = \begin{cases} 1, & \text{if } x \in \{X \geq t\} \\ 0, & \text{otherwise.} \end{cases}$$

Since ψ is a positive, non-decreasing function, we have

$$\begin{aligned} \mathrm{P}(X \geq t) &= 0 \cdot \mathrm{P}(X < t) + 1 \cdot \mathrm{P}(X \geq t) \\ &= \mathrm{E}\left(1_{\{X \geq t\}}\right) \\ &\leq \mathrm{E}\left(\frac{\psi(X)}{\psi(t)} 1_{\{X \geq t\}}\right) \\ &= \frac{1}{\psi(t)} \mathrm{E}\left(\psi(X) \cdot 1_{\{X \geq t\}}\right) \\ &\leq \end{aligned}$$

Proof of the Markov Inequality

Let $1_{\{X \geq t\}}$ be the indicator function of the set/event $\{X \geq t\}$. I.e.

$$1_{\{X \geq t\}}(x) = \begin{cases} 1, & \text{if } x \in \{X \geq t\} \\ 0, & \text{otherwise.} \end{cases}$$

Since ψ is a positive, non-decreasing function, we have

$$\begin{aligned} \mathrm{P}(X \geq t) &= 0 \cdot \mathrm{P}(X < t) + 1 \cdot \mathrm{P}(X \geq t) \\ &= \mathrm{E}\left(1_{\{X \geq t\}}\right) \\ &\leq \mathrm{E}\left(\frac{\psi(X)}{\psi(t)} 1_{\{X \geq t\}}\right) \\ &= \frac{1}{\psi(t)} \mathrm{E}\left(\psi(X) \cdot 1_{\{X \geq t\}}\right) \\ &\leq \frac{1}{\psi(t)} \mathrm{E}\left(\psi(X)\right). \end{aligned}$$

Proof of the Markov Inequality

Let $1_{\{X \geq t\}}$ be the indicator function of the set/event $\{X \geq t\}$. I.e.

$$1_{\{X \geq t\}}(x) = \begin{cases} 1, & \text{if } x \in \{X \geq t\} \\ 0, & \text{otherwise.} \end{cases}$$

Since ψ is a positive, non-decreasing function, we have

$$\begin{aligned} \mathrm{P}(X \geq t) &= 0 \cdot \mathrm{P}(X < t) + 1 \cdot \mathrm{P}(X \geq t) \\ &= \mathrm{E}\left(1_{\{X \geq t\}}\right) \\ &\leq \mathrm{E}\left(\frac{\psi(X)}{\psi(t)} 1_{\{X \geq t\}}\right) \\ &= \frac{1}{\psi(t)} \mathrm{E}\left(\psi(X) \cdot 1_{\{X \geq t\}}\right) \\ &\leq \frac{1}{\psi(t)} \mathrm{E}\left(\psi(X)\right). \end{aligned}$$

$\square$

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:

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

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

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

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:

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:

$$\mathrm{P}\left(X \in B_{p,\delta}\right) \quad =$$

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:

$$\begin{aligned} \mathrm{P}\left(X \in B_{p,\delta}\right) &= \mathrm{P}\left(\mathrm{e}^{-||X||^2/2} \geq \delta\right) \\ &\leq \end{aligned}$$

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:

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

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:

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

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:

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

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

So for p large we see that most of the mass of the standard Gaussian distribution is in the tail.

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

So for p large we see that most of the mass of the standard Gaussian distribution is in the tail. If we want to have $\mathrm{P}(X \in B_{p,\delta}) \geq 1/2$, we must have $\delta \leq 2^{-p/2+1}$ ...

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

So for p large we see that most of the mass of the standard Gaussian distribution is in the tail. If we want to have $\mathrm{P}(X \in B_{p,\delta}) \geq 1/2$, we must have $\delta \leq 2^{-p/2+1}$ ... which is exponentially small.

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

So for p large we see that most of the mass of the standard Gaussian distribution is in the tail. If we want to have $\mathrm{P}(X \in B_{p,\delta}) \geq 1/2$, we must have $\delta \leq 2^{-p/2+1}$ ... which is exponentially small.

How to understand this counter-intuitive phenomenon?

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

So for p large we see that most of the mass of the standard Gaussian distribution is in the tail. If we want to have $\mathrm{P}(X \in B_{p,\delta}) \geq 1/2$, we must have $\delta \leq 2^{-p/2+1}$ ... which is exponentially small.

How to understand this counter-intuitive phenomenon? It all has to do with the geometric properties of high-dimensional spaces described earlier;

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

So for p large we see that most of the mass of the standard Gaussian distribution is in the tail. If we want to have $\mathrm{P}(X \in B_{p,\delta}) \geq 1/2$, we must have $\delta \leq 2^{-p/2+1}$ ... which is exponentially small.

How to understand this counter-intuitive phenomenon? It all has to do with the geometric properties of high-dimensional spaces described earlier; with p large there's a vast space out there which need to be filled with mass.

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*.

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 ...

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!

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!

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!

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.

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.

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.

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.

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:

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

- pixel intensities in images are not purely random, since many geometrical structures exist in the images.

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

- pixel intensities in images are not purely random, since many geometrical structures exist in the images.
- biological data are the outcome of a highly regulated biological system, so the data has a relatively low complexity.

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

- pixel intensities in images are not purely random, since many geometrical structures exist in the images.
- biological data are the outcome of a highly regulated biological system, so the data has a relatively low complexity.
- marketing data reflects some social structure in society; these structures are relatively simple.

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

- pixel intensities in images are not purely random, since many geometrical structures exist in the images.
- biological data are the outcome of a highly regulated biological system, so the data has a relatively low complexity.
- marketing data reflects some social structure in society; these structures are relatively simple.
- technical data are the outcome of human technologies, and therefore limited in their complexity.

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

- pixel intensities in images are not purely random, since many geometrical structures exist in the images.
- biological data are the outcome of a highly regulated biological system, so the data has a relatively low complexity.
- marketing data reflects some social structure in society; these structures are relatively simple.
- technical data are the outcome of human technologies, and therefore limited in their complexity.

So in many case the data have intrinsic low complexity,

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

- pixel intensities in images are not purely random, since many geometrical structures exist in the images.
- biological data are the outcome of a highly regulated biological system, so the data has a relatively low complexity.
- marketing data reflects some social structure in society; these structures are relatively simple.
- technical data are the outcome of human technologies, and therefore limited in their complexity.

So in many case the data have intrinsic low complexity, and we can try extract useful information from them once we can localize these lower dimensional structures

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

- pixel intensities in images are not purely random, since many geometrical structures exist in the images.
- biological data are the outcome of a highly regulated biological system, so the data has a relatively low complexity.
- marketing data reflects some social structure in society; these structures are relatively simple.
- technical data are the outcome of human technologies, and therefore limited in their complexity.

So in many case the data have intrinsic low complexity, and we can try extract useful information from them once we can localize these lower dimensional structures (and then use our lower-dimensional statistics to them).

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

- pixel intensities in images are not purely random, since many geometrical structures exist in the images.
- biological data are the outcome of a highly regulated biological system, so the data has a relatively low complexity.
- marketing data reflects some social structure in society; these structures are relatively simple.
- technical data are the outcome of human technologies, and therefore limited in their complexity.

So in many case the data have intrinsic low complexity, and we can try extract useful information from them once we can localize these lower dimensional structures (and then use our lower-dimensional statistics to them).

The problem of course is, that these lower-dimensional structures are *unknown*;

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

- pixel intensities in images are not purely random, since many geometrical structures exist in the images.
- biological data are the outcome of a highly regulated biological system, so the data has a relatively low complexity.
- marketing data reflects some social structure in society; these structures are relatively simple.
- technical data are the outcome of human technologies, and therefore limited in their complexity.

So in many case the data have intrinsic low complexity, and we can try extract useful information from them once we can localize these lower dimensional structures (and then use our lower-dimensional statistics to them).

The problem of course is, that these lower-dimensional structures are *unknown*; the main task is to identify at least approximately these structures.

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

- pixel intensities in images are not purely random, since many geometrical structures exist in the images.
- biological data are the outcome of a highly regulated biological system, so the data has a relatively low complexity.
- marketing data reflects some social structure in society; these structures are relatively simple.
- technical data are the outcome of human technologies, and therefore limited in their complexity.

So in many case the data have intrinsic low complexity, and we can try extract useful information from them once we can localize these lower dimensional structures (and then use our lower-dimensional statistics to them).

The problem of course is, that these lower-dimensional structures are *unknown*; the main task is to identify at least approximately these structures. According to Giraud this is the central issue of high-dimensional statistics.

A Paradigm Shift

In fact, in his introductory chapter Giraud claims that a *Paradigm Shift* is needed.

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 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 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

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 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 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

The classical analysis in which we assume p fixed and n tending to infinity doesn't always seem to make sense anymore.

A Paradigm Shift

The classical analysis in which we assume p fixed and n tending to infinity doesn't always seem to make sense anymore. Giraud thinks it is even worse:

A Paradigm Shift

The classical analysis in which we assume p fixed and n tending to infinity doesn't always seem to make sense anymore. Giraud thinks it is even worse: the classical approach might lead to misleading or even wrong conclusions!

A Paradigm Shift

The classical analysis in which we assume p fixed and n tending to infinity doesn't always seem to make sense anymore. Giraud thinks it is even worse: the classical approach might lead to misleading or even wrong conclusions!

According to Giraud we must change our point of view of statistics!

A Paradigm Shift

The classical analysis in which we assume p fixed and n tending to infinity doesn't always seem to make sense anymore. Giraud thinks it is even worse: the classical approach might lead to misleading or even wrong conclusions!

According to Giraud we must change our point of view of statistics!

Giraud's point of view is not entirely new.

A Paradigm Shift

The classical analysis in which we assume p fixed and n tending to infinity doesn't always seem to make sense anymore. Giraud thinks it is even worse: the classical approach might lead to misleading or even wrong conclusions!

According to Giraud we must change our point of view of statistics!

Giraud's point of view is not entirely new. Already in 2000 David Donaho wrote:

A Paradigm Shift

The classical analysis in which we assume p fixed and n tending to infinity doesn't always seem to make sense anymore. Giraud thinks it is even worse: the classical approach might lead to misleading or even wrong conclusions!

According to Giraud we must change our point of view of statistics!

Giraud's point of view is not entirely new. Already in 2000 David Donaho wrote:

> "Classical methods are simply not designed to cope with this kind of explosive growth of dimensionality of the observation vector. We can say with complete condence that in the coming century, high-dimensional data analysis will be a very signicant activity, and completely new methods of high-dimensional data analysis will be developed; we just dont know what they are yet."

A Paradigm Shift

One of the possible approaches (the one Giraud advocates)

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.

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)

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;

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

To be able to quantify the performance of an estimator in a non-asymptotic way we need to "replace" the classical tools like the *Central Limit Theorem* and the *Law of Large Numbers* by non-asysmtotic results.

Mathematics of High-Dimensional Statistics

To be able to quantify the performance of an estimator in a non-asymptotic way we need to “replace” the classical tools like the *Central Limit Theorem* and the *Law of Large Numbers* by non-asysmtotic results.

Recall for example the CLT;

Mathematics of High-Dimensional Statistics

To be able to quantify the performance of an estimator in a non-asymptotic way we need to "replace" the classical tools like the *Central Limit Theorem* and the *Law of Large Numbers* by non-asysmtotic results.

Recall for example the CLT; if $f : \mathbb{R} \to \mathbb{R}$ is a function and $X_1, X_2, \ldots, X_n$ are i.i.d. random variables,

Mathematics of High-Dimensional Statistics

To be able to quantify the performance of an estimator in a non-asymptotic way we need to "replace" the classical tools like the *Central Limit Theorem* and the *Law of Large Numbers* by non-asysmtotic results.

Recall for example the CLT; if $f : \mathbb{R} \to \mathbb{R}$ is a function and $X_1, X_2, \ldots, X_n$ are i.i.d. random variables, such that $\text{Var}(f(X_1)) < \infty$,

Mathematics of High-Dimensional Statistics

To be able to quantify the performance of an estimator in a non-asymptotic way we need to "replace" the classical tools like the *Central Limit Theorem* and the *Law of Large Numbers* by non-asysmtotic results.

Recall for example the CLT; if $f : \mathbb{R} \to \mathbb{R}$ is a function and $X_1, X_2, \ldots, X_n$ are i.i.d. random variables, such that $\text{Var}(f(X_1)) < \infty$, then we have when $n \to \infty$ that:

Mathematics of High-Dimensional Statistics

To be able to quantify the performance of an estimator in a non-asymptotic way we need to "replace" the classical tools like the *Central Limit Theorem* and the *Law of Large Numbers* by non-asysmtotic results.

Recall for example the CLT; if $f : \mathbb{R} \to \mathbb{R}$ is a function and $X_1, X_2, \ldots, X_n$ are i.i.d. random variables, such that $\mathrm{Var}(f(X_1)) < \infty$, then we have when $n \to \infty$ that:

$$\sqrt{\frac{n}{\mathrm{Var}(f(X_1))}} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_1) - \mathrm{E}(f(X_1)) \right) \to Z \qquad \text{(in distribution)},$$

Mathematics of High-Dimensional Statistics

To be able to quantify the performance of an estimator in a non-asymptotic way we need to "replace" the classical tools like the *Central Limit Theorem* and the *Law of Large Numbers* by non-asysmtotic results.

Recall for example the CLT; if $f : \mathbb{R} \to \mathbb{R}$ is a function and $X_1, X_2, \ldots, X_n$ are i.i.d. random variables, such that $\text{Var}(f(X_1)) < \infty$, then we have when $n \to \infty$ that:

$$\sqrt{\frac{n}{\text{Var}(f(X_1))}} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_1) - \mathrm{E}(f(X_1)) \right) \to Z \qquad \text{(in distribution)},$$

where Z is a random variable with a standard normal distribution.

Mathematics of High-Dimensional Statistics

To be able to quantify the performance of an estimator in a non-asymptotic way we need to "replace" the classical tools like the *Central Limit Theorem* and the *Law of Large Numbers* by non-asysmtotic results.

Recall for example the CLT; if $f : \mathbb{R} \to \mathbb{R}$ is a function and $X_1, X_2, \ldots, X_n$ are i.i.d. random variables, such that $\text{Var}(f(X_1)) < \infty$, then we have when $n \to \infty$ that:

$$\sqrt{\frac{n}{\text{Var}(f(X_1))}} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_1) - \mathrm{E}(f(X_1)) \right) \to Z \qquad \text{(in distribution)},$$

where Z is a random variable with a standard normal distribution.

If we moreover assume that f is L-Lipschitz, and X_1 and X_2 i.i.d. with finite variance $\sigma^2(> 0)$, then

Mathematics of High-Dimensional Statistics

To be able to quantify the performance of an estimator in a non-asymptotic way we need to "replace" the classical tools like the *Central Limit Theorem* and the *Law of Large Numbers* by non-asysmtotic results.

Recall for example the CLT; if $f : \mathbb{R} \to \mathbb{R}$ is a function and $X_1, X_2, \dots, X_n$ are i.i.d. random variables, such that $\text{Var}(f(X_1)) < \infty$, then we have when $n \to \infty$ that:

$$\sqrt{\frac{n}{\text{Var}(f(X_1))}} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_1) - \mathrm{E}(f(X_1)) \right) \to Z \qquad \text{(in distribution)},$$

where Z is a random variable with a standard normal distribution.

If we moreover assume that f is L-Lipschitz, and X_1 and X_2 i.i.d. with finite variance $\sigma^2 (> 0)$, then

$$\text{Var}(f(X_1)) =$$

Mathematics of High-Dimensional Statistics

To be able to quantify the performance of an estimator in a non-asymptotic way we need to "replace" the classical tools like the *Central Limit Theorem* and the *Law of Large Numbers* by non-asysmtotic results.

Recall for example the CLT; if $f : \mathbb{R} \to \mathbb{R}$ is a function and $X_1, X_2, \ldots, X_n$ are i.i.d. random variables, such that $\mathsf{Var}(f(X_1)) < \infty$, then we have when $n \to \infty$ that:

$$\sqrt{\frac{n}{\mathsf{Var}(f(X_1))}} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_1) - \mathrm{E}(f(X_1)) \right) \to Z \qquad \text{(in distribution)},$$

where Z is a random variable with a standard normal distribution.

If we moreover assume that f is L-Lipschitz, and X_1 and X_2 i.i.d. with finite variance $\sigma^2 (> 0)$, then

$$\mathsf{Var}(f(X_1)) = \frac{1}{2} \mathrm{E}\left((f(X_1) - f(X_2))^2 \right) \leq$$

Mathematics of High-Dimensional Statistics

To be able to quantify the performance of an estimator in a non-asymptotic way we need to "replace" the classical tools like the *Central Limit Theorem* and the *Law of Large Numbers* by non-asysmtotic results.

Recall for example the CLT; if $f : \mathbb{R} \to \mathbb{R}$ is a function and $X_1, X_2, \ldots, X_n$ are i.i.d. random variables, such that $\text{Var}(f(X_1)) < \infty$, then we have when $n \to \infty$ that:

$$\sqrt{\frac{n}{\text{Var}(f(X_1))}} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_1) - \mathrm{E}(f(X_1)) \right) \to Z \qquad \text{(in distribution)},$$

where Z is a random variable with a standard normal distribution.

If we moreover assume that f is L-Lipschitz, and X_1 and X_2 i.i.d. with finite variance $\sigma^2 (> 0)$, then

$$\text{Var}(f(X_1)) = \frac{1}{2} \mathrm{E}\left((f(X_1) - f(X_2))^2 \right) \leq \frac{L^2}{2} \mathrm{E}\left((X_1 - X_2)^2 \right) =$$

Mathematics of High-Dimensional Statistics

To be able to quantify the performance of an estimator in a non-asymptotic way we need to "replace" the classical tools like the *Central Limit Theorem* and the *Law of Large Numbers* by non-asysmtotic results.

Recall for example the CLT; if $f : \mathbb{R} \to \mathbb{R}$ is a function and $X_1, X_2, \ldots, X_n$ are i.i.d. random variables, such that $\text{Var}(f(X_1)) < \infty$, then we have when $n \to \infty$ that:

$$\sqrt{\frac{n}{\text{Var}(f(X_1))}} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_1) - \mathrm{E}(f(X_1)) \right) \to Z \qquad \text{(in distribution)},$$

where Z is a random variable with a standard normal distribution.

If we moreover assume that f is L-Lipschitz, and X_1 and X_2 i.i.d. with finite variance $\sigma^2 (> 0)$, then

$$\text{Var}(f(X_1)) = \frac{1}{2}\mathrm{E}\left((f(X_1) - f(X_2))^2\right) \leq \frac{L^2}{2}\mathrm{E}\left((X_1 - X_2)^2\right) = L^2\sigma^2. \tag{4}$$

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:

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:

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

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:

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

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 $x > 0$.

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.

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.

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

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 $X_1, X_2, \dots, X_n$ 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

From this Concentration Inequality we derive, if $X_1, \dots, X_n$ are i.i.d., with $N(0, \sigma^2)$ Gaussian distributions, that

Mathematics of High-Dimensional Statistics

From this Concentration Inequality we derive, if $X_1, \ldots, X_n$ are i.i.d., with $N(0, \sigma^2)$ Gaussian distributions, that

$$F(X_1, \ldots, X_n) - \mathrm{E}(F(X_1, \ldots, X_n)) \leq$$

Mathematics of High-Dimensional Statistics

From this Concentration Inequality we derive, if $X_1, \ldots, X_n$ are i.i.d., with $N(0, \sigma^2)$ Gaussian distributions, that

$$F(X_1, \ldots, X_n) - \mathrm{E}(F(X_1, \ldots, X_n)) \leq L\sigma\sqrt{2\xi},$$

Mathematics of High-Dimensional Statistics

From this Concentration Inequality we derive, if $X_1, \ldots, X_n$ are i.i.d., with $N(0, \sigma^2)$ Gaussian distributions, that

$$F(X_1, \ldots, X_n) - \mathrm{E}(F(X_1, \ldots, X_n)) \leq L\sigma\sqrt{2\xi},$$

where $\mathrm{P}(\xi \geq t) \leq \mathrm{e}^{-t}$ for $t \geq 0$.

Mathematics of High-Dimensional Statistics

From this Concentration Inequality we derive, if $X_1, \ldots, X_n$ are i.i.d., with $N(0, \sigma^2)$ Gaussian distributions, that

$$F(X_1, \ldots, X_n) - \mathrm{E}(F(X_1, \ldots, X_n)) \leq L\sigma\sqrt{2\xi},$$

where $\mathrm{P}(\xi \geq t) \leq \mathrm{e}^{-t}$ for $t \geq 0$.

When $f : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz, the *Cauchy-Schwatz*-inequality yields

Mathematics of High-Dimensional Statistics

From this Concentration Inequality we derive, if $X_1, \ldots, X_n$ are i.i.d., with $N(0, \sigma^2)$ Gaussian distributions, that

$$F(X_1, \ldots, X_n) - \mathrm{E}(F(X_1, \ldots, X_n)) \leq L\sigma\sqrt{2\xi},$$

where $\mathrm{P}(\xi \geq t) \leq \mathrm{e}^{-t}$ for $t \geq 0$.

When $f : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz, the *Cauchy-Schwatz*-inequality yields

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \frac{1}{n} \sum_{i=1}^{n} f(Y_i) \right| \leq$$

Mathematics of High-Dimensional Statistics

From this Concentration Inequality we derive, if $X_1, \ldots, X_n$ are i.i.d., with $N(0, \sigma^2)$ Gaussian distributions, that

$$F(X_1, \ldots, X_n) - \mathrm{E}(F(X_1, \ldots, X_n)) \leq L\sigma\sqrt{2\xi},$$

where $\mathrm{P}(\xi \geq t) \leq \mathrm{e}^{-t}$ for $t \geq 0$.

When $f : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz, the *Cauchy-Schwatz*-inequality yields

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \frac{1}{n} \sum_{i=1}^{n} f(Y_i) \right| \leq \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - f(Y_i)| \leq$$

Mathematics of High-Dimensional Statistics

From this Concentration Inequality we derive, if $X_1, \ldots, X_n$ are i.i.d., with $N(0, \sigma^2)$ Gaussian distributions, that

$$F(X_1, \ldots, X_n) - \mathrm{E}(F(X_1, \ldots, X_n)) \leq L\sigma\sqrt{2\xi},$$

where $\mathrm{P}(\xi \geq t) \leq \mathrm{e}^{-t}$ for $t \geq 0$.

When $f : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz, the *Cauchy-Schwatz*-inequality yields

$$\left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \frac{1}{n}\sum_{i=1}^{n} f(Y_i) \right| \leq \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - f(Y_i)| \leq \frac{L}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2},$$

Mathematics of High-Dimensional Statistics

From this Concentration Inequality we derive, if $X_1, \ldots, X_n$ are i.i.d., with $N(0, \sigma^2)$ Gaussian distributions, that

$$F(X_1, \ldots, X_n) - \mathrm{E}(F(X_1, \ldots, X_n)) \leq L\sigma\sqrt{2\xi},$$

where $\mathrm{P}(\xi \geq t) \leq \mathrm{e}^{-t}$ for $t \geq 0$.

When $f : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz, the *Cauchy-Schwatz*-inequality yields

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \frac{1}{n} \sum_{i=1}^{n} f(Y_i) \right| \leq \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - f(Y_i)| \leq \frac{L}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2},$$

so the function

Mathematics of High-Dimensional Statistics

From this Concentration Inequality we derive, if $X_1, \ldots, X_n$ are i.i.d., with $N(0, \sigma^2)$ Gaussian distributions, that

$$F(X_1, \ldots, X_n) - \mathrm{E}(F(X_1, \ldots, X_n)) \leq L\sigma\sqrt{2\xi},$$

where $\mathrm{P}(\xi \geq t) \leq \mathrm{e}^{-t}$ for $t \geq 0$.

When $f : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz, the *Cauchy-Schwatz*-inequality yields

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \frac{1}{n} \sum_{i=1}^{n} f(Y_i) \right| \leq \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - f(Y_i)| \leq \frac{L}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2},$$

so the function

$$F(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

Mathematics of High-Dimensional Statistics

From this Concentration Inequality we derive, if $X_1, \ldots, X_n$ are i.i.d., with $N(0, \sigma^2)$ Gaussian distributions, that

$$F(X_1, \ldots, X_n) - \mathrm{E}(F(X_1, \ldots, X_n)) \leq L\sigma\sqrt{2\xi},$$

where $\mathrm{P}(\xi \geq t) \leq \mathrm{e}^{-t}$ for $t \geq 0$.

When $f : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz, the *Cauchy-Schwatz*-inequality yields

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \frac{1}{n} \sum_{i=1}^{n} f(Y_i) \right| \leq \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - f(Y_i)| \leq \frac{L}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2},$$

so the function

$$F(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

is $n^{-1/2}L$-Lipschitz.

Mathematics of High-Dimensional Statistics

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

Mathematics of High-Dimensional Statistics

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

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

Mathematics of High-Dimensional Statistics

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

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

Mathematics of High-Dimensional Statistics

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

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

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 (4).

References

David L. Donaho – *High-dimensional data analysis: The cursus and blessings of dimensionality*. 200, American Mathematical Society "Math Challenges of the 21st Century."

Christophe Giraud – *Introduction to High-Dimensional Statistics*, CRC Press, Monographs on Statistics and Applied Probability 139, Boca Raton, 2015.