Lecture 16: Concentration of Measure Thin Tails in Big Data

“The idea of concentration of measure, which was discovered by Vitali Milman, is arguably one of the great ideas of analysis in our times” (M. Talagrand 1996)

The class lectures have gone back and forth between, on the one side, an effort to show how power laws that characterize extreme events arise in the world from statistical processes and economic behavior and on the other side, examples of the great importance of rare discontinuous events on economic life. From this perspective, big data is critical to identifying and estimating power laws and thus better understanding the few important situations that shake up the world. Since we need large # of observations to find out what happens in the fat tail, we need big data. At the same time, to the extent that the processes that generate outcomes for most observations work at the tails, the more we learn from the non-tailed events, the better adept we will be at analyzing the tail events.

To estimate income of billionaires, need big data on incomes to get enough observations to measure rare billionaires' income. The more billionaires income is determined by similar processes as the income of the rest of us, the more likely will the observations of the many cast predictive/other light on the “rare” event,

Concentration of measure shifts the perspective on big data from helping us to understand infrequent tailed events to determine better what is normal/average. It shifts focus from big data giving *many observations about rare* events or variables to big data providing *observations of many events* or variables. Instead of fat tails, it makes thin tails the “hero” of the story. Three parts to topic: 1) Introduce to two forms of big data; 2) Explain concentration of measure phenomenon and inequalities; 3) Relate statistical learning data-mining models to concentration of measure.

1, BIG DATA IS Number observations x Number of Variables

(Getting a Big Data Job for Dummies has Four Vs: Volume, Variety, Veracity, Velocity_

Data-mining is about large data sets, but how big is BIG and what has to be big to get insight?

Economics mostly focuses on millions of observations on a limited number of variables – a Census of people in a given area that dominates survey samples; data on output, inputs, patents, investment of tens or hundreds of thousands of firms over 20-30 years. Large N with limited number of variables.

Medical practice/research/clinical studies often about a small number of people – just the patient --- with data on their many attributes – the human genome – detailed observations on the neurons in a mouse's brain as it “plays” a virtual reality game; medical measurements of someone with rare disease.

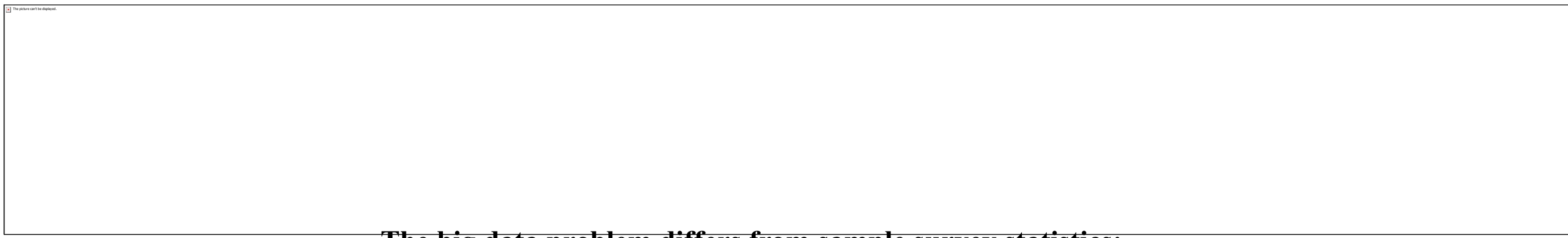
Google cloud big data sets publicly available: Cloud Genomics The 1000 Genomes dataset comprises roughly 2,500 genomes from 25 populations around the world. See the 1000 Genomes Project website Pilot publication: An integrated map of genetic variation from 1,092 human genomes. Lots of information on a few.

shakespeare giving the number of times each word appears in each corpus; lots of words but just one measure

trigrams Contains English language trigrams from a sample of works published between 1520 and 2008.

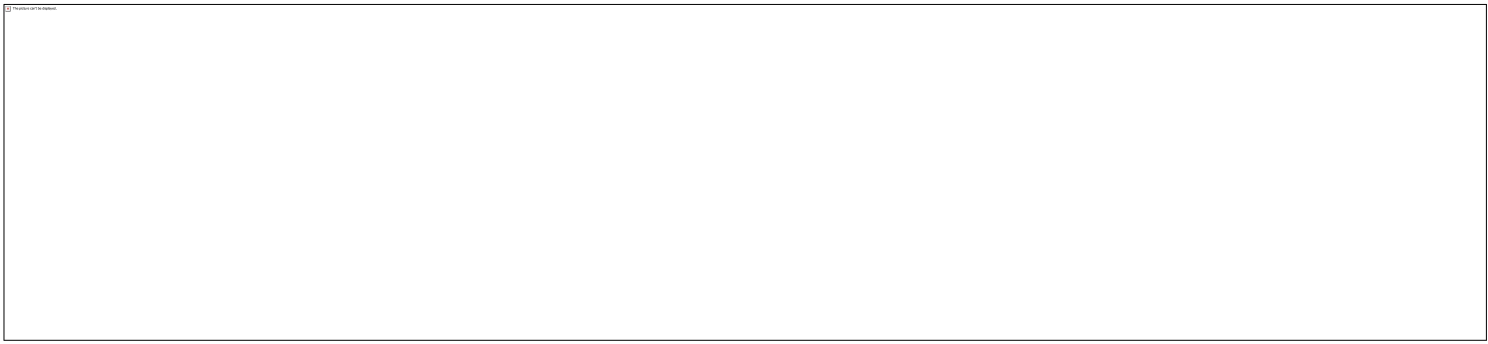
wikipedia Contains the complete revision history for all Wikipedia articles up to April 2010.

The defining feature of big data is that # observations (N) x # of variables (V) is large. Direction in which it is large will determine what one can do with it. Take transactions data.



The big data problem differs from sample survey statistics:

Standard stats, which assume we know the right model and need to estimate parameters and standard errors. Here the goal is to find the right model/function from a space of possible models (functions).



Small N and large numbers of variables

Say you want to predict who will go out with whom after a social with 20 people? (event before Coronavirus limited human interactions)

One way to predict would be to ask people who they spoke to and for how long; mutual interests or friends. You might gather information on 3-4 variables and try some regression type model that would improve as you increased the sample size. But how about making a movie of the event. Now we have much more information. Maybe the best predictor is # of smiles vs # frowns in an interaction; but maybe a wink from someone who is not talking with you is more likely to produce a date? Or maybe someone who listens to 2-3 people and asks for their email gets the date. With a good enough camera we might capture flushes of the cheek or some other body language. Or maybe we have some "unobtrusive" brain scan machine or bio-marker. In this approach we might have dozens of measures on each the 20 people that could predict outcomes. But how do you analyze data with more variables than observations?

Classical statistics is about V variables with N—> infinity; you have limit theorems as N gets larger with inequalities.

Flip a coin once and chance of head is ½ but you never get ½ either get head or tails. Flip it 1,000s of times and get average of 0.50 or so with increasingly narrow band around average as N gets larger. Law of large numbers.

New data observations is about N fixed and V—> infinity.

In classical world, dimensionality is a curse. — too many variables requires huge samples to estimate the effect of the variables at reasonable level of error . But in concentration of measure world dimensionality is a virtue.

Donoho (Stanford Stats, 2000, https://pdfs.semanticscholar.org/63c6/8278418b69f60b4814fae8dd15b1b1854295.pdf)
"The trend today is towards more observations but even more so, to **radically larger numbers of variables** – … hyper-informative detail about each observed instance. … observations gathered on individual instances are curves, or spectra, or images, or even movies, so that a single observation has dimensions in the thousands or billions, while there are only tens or hundreds of instances available for study. Classical methods are simply not designed to cope with this kind of explosive growth of dimensionality of the observation vector...high-dimensional function.

The blessings of dimensionality are less widely noted, but they include the **concentration of measure** phenomenon (so-called in the geometry of Banach spaces), which means that certain random fluctuations are very well controlled in high dimensions … which suggest that statements about very high-dimensional settings may be made where moderate dimensions would be too complicated ….

2.What is concentration of measure?

V.Milman theorem about probabilities on product spaces in high dimensions. Consider a function f on the d-dimensional sphere that is smooth so that a derivative in any dimension never gets too big. This is Lispschitz condition). In this situation the **probability concentrates around average** so that chance you are far away from the average drops sharply.

The picture can't be displayed.

Wikipedia: "**A random variable that depends in a Lipschitz way on many independent variables (but not too much on any of them) is essentially constant**". Prob that f(x) differs from E(f(x) differs from the mean by less than t depends on exp-ct^2 ,where c is some constant.

Problem goes back to Queen Dido of Carthage! – isoperimetric inequality. Place a uniform measure P on the sphere, and let X be a random variable distributed P Then P{|f(x) − Ef(x)| > t} ≤ C1 exp(−C2t 2).

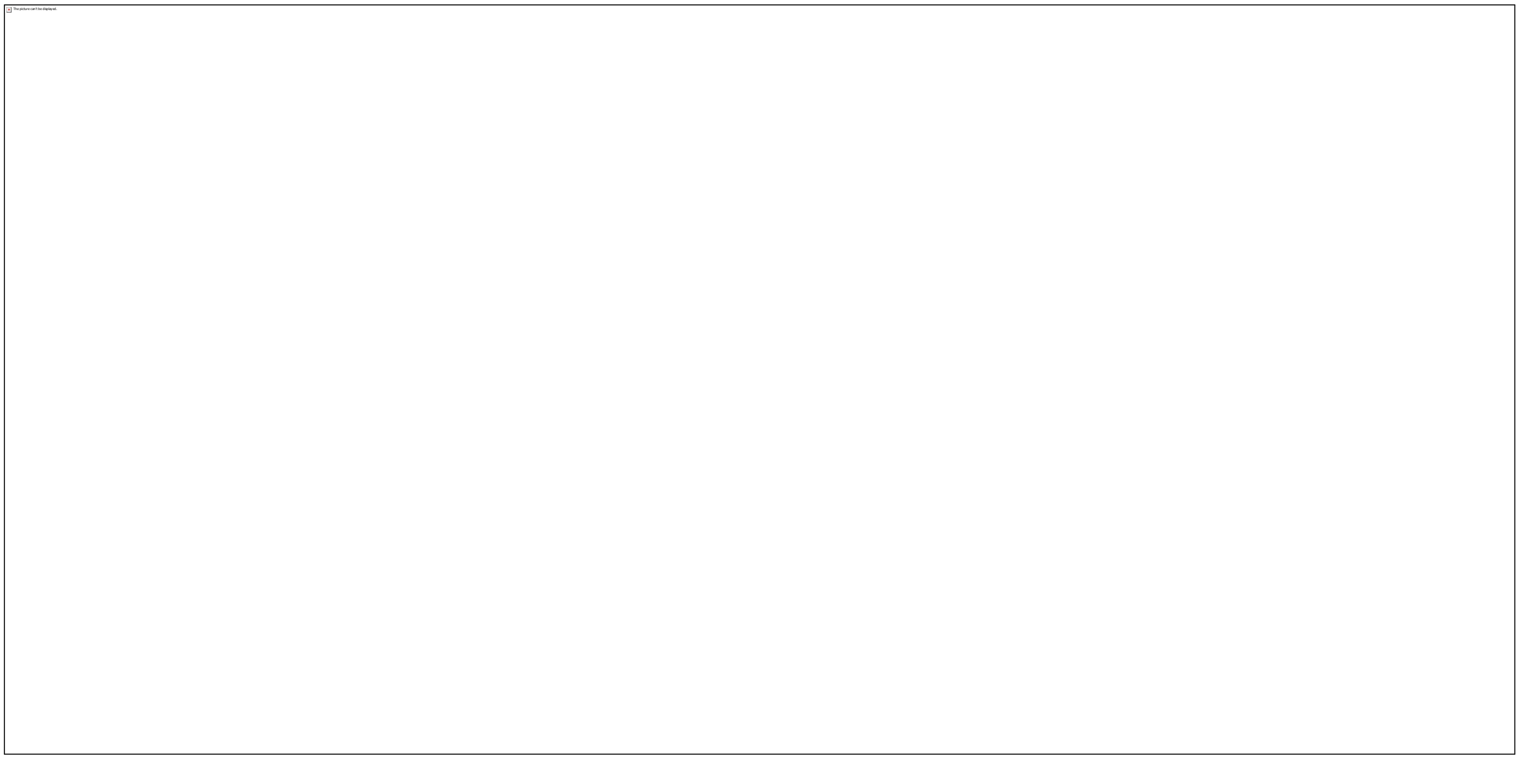
In standard stats closest parallel is use of principal components to identify pattern in multidimensional data

The picture can't be displayed.

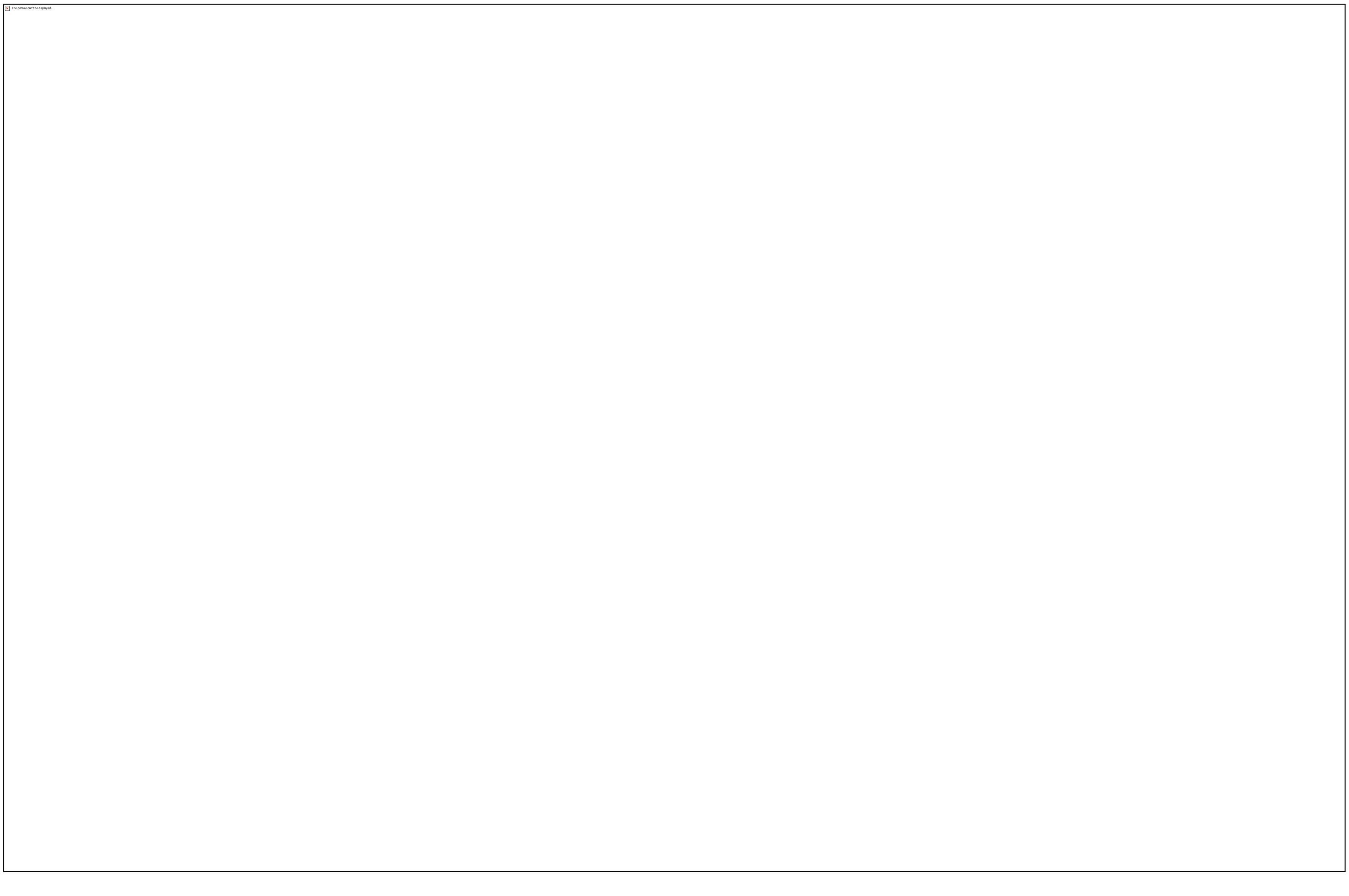
The picture can't be displayed.

The picture can't be displayed.

An Example: GLOBAL INNOVATION INDEX https://www.globalinnovationindex.org/gii-2018-report
Business groups produce lots of indices of economic attributes where we lack a single well-defined measure such as GDP and use INDICATORS



How do you make an indicator of innovation? No official statistics on innovative activity so Innovation index uses 82 *indicators* on the basis of literature review, expert opinion, country coverage, and timeliness, which fall into three categories: quantitative/objective/hard data (58 indicators), composite indicators/index data (19 indicators), and survey/ qualitative/subjective/ soft data (5 indicators)**. O**verall score is simple average of Input and Output Sub-Index scores.



- The Innovation Input SubIndex is comprised of five input pillars that capture elements of the national economy that enable innovative activities: (1) Institutions, (2) Human capital and research, (3) Infrastructure, (4) Market sophistication, and (5) Business sophistication.
- The Innovation Output SubIndex provides information about outputs that are the results of innovative activities within the economy. There are two output pillars: (6) Knowledge and technology outputs and (7) Creative outputs.
- The Innovation Efficiency Ratio is the ratio of the Output Sub-Index score over the Input Sub-Index score. It shows how much innovation output a given country is getting for its inputs.

The weighting is semi-arbitrary but with enough indicators this may not matter due to **"concentration of measure"** phenomenon. Instead of fat tails, concentration of measure lives on thin tails, where **the mass in the tails decays rapidly – exponentially.** Observed possibilities concentrate in a narrow range close to what we want to estimate. With lots of **independent** dimensions/measures **any** combination (say average of the measures) concentrates in a narrow space around the mean value. If the variables are independent it is nearly impossible for them to "work together" to simultaneously pull the average far from its mean. **Independence is the key;** concentration of measure results fail if variables are highly correlated with each other.

Just as s**tructure among variables gives fat tails, independence gives thin tails**. This is "central limit theorem" writ large.

The global innovation folk evaluate their measure via PCA."PCA results confirm the presence of a single latent dimension in each of the seven pillars (one component with an eigenvalue greater than 1.0) that captures between close to 60% (pillar 4: Market sophistication) up to 82% (pillar 1: Institutions) of the total variance in the three underlying sub-pillars. Furthermore,results confirm the expectation that the subpillars are more correlated with their own pillar than with any other pillar and that all correlation coefficients are close to or greater than 0.70. The five input pillars share a single statistical dimension that summarizes 82% of the total variance, and the five loadings (correlation and sub-pillars can explain a similar amount of variance in their respective sub-pillars /pillars)."

3. **Statistical learning theory/machine learning/data-mining**

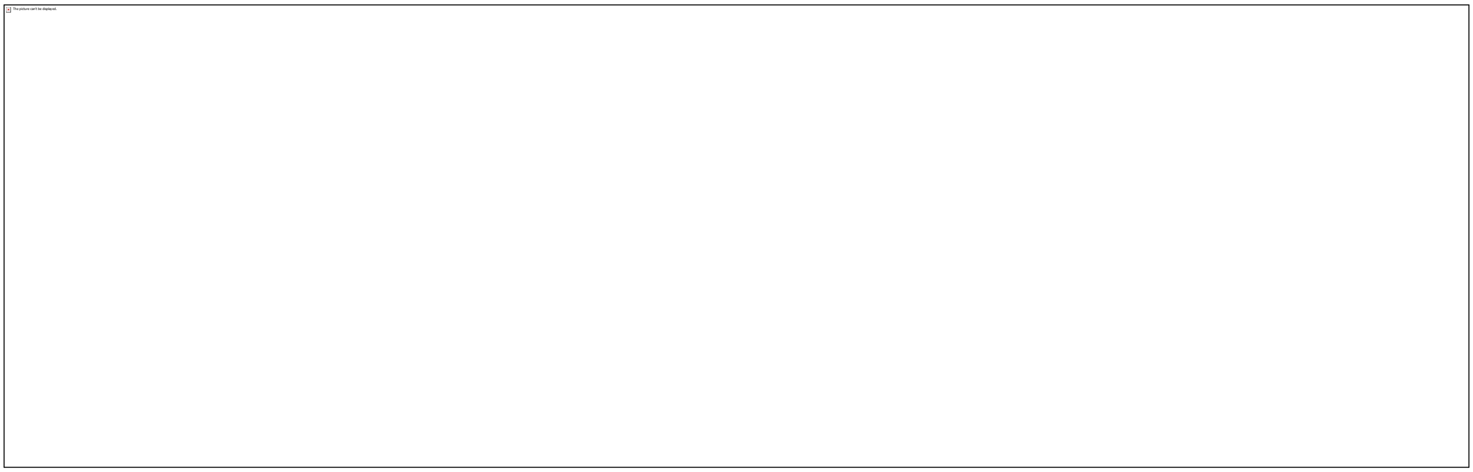
Statistical learning theory shows that we can reach valid conclusions about the population from even a non-representative sample under the "innocuous" assumption of smoothness in the data x variable space.

Consider a non-random but large web-based survey. Under what conditions will estimates of a model from a non-representative sample tells us something useful about the entire population. The trick is to generalize from the estimates of the model and its error on existing data (which is all we observe) and on measures of the complexity of the *model* (which we control) to an estimate of how that model fits unseen data.

This type of generalization follows the logic "fit on the training data" but "decide on what works from the test data set". That smoothness is necessary is mindful of Kaufman's NK landscape model: with a correlated landscape you learn something about payoff in nearby points but you learn nothing if the landscape is random/very rugged.

Here is one way to see the statistical learning theory problem. We have 7 data points linking y=f(x) to x. Many functions fit these data points. The two R(p) functions fit the data perfectly. Statistical learning theory says that with reasonably innocuous assumptions, you can find "the best function" for relating y to x over the entire space!

What might you do?



Scheme is to specify the space of models to use (polynomials, sin/cosine Fourier transforms, or both or ... etc). We want to find the function that gives the best result — say smallest squared error — over the entire space.

We have observed data and **assume some sort of smoothness** — that the data are not perversely chosen even though they cover only part of the space. Think training set (observations) and test set (other data), and generalizing to the rest of the data. The smoothness assumption is the general functional form of the relation; linear, polynomial, neural net with one hidden layer, CART model ,where your analysis determines which works best.

Say you limit your model to polynomials and assume that a real polynomial R(x) fits the data everywhere. You find an approximation A(x). The goal is to get an approximation that gives the best expected performance over the ENTIRE space, expected because you do not observe the entire space. You just observe data set Z. The way to find A is to specify a loss function L(data set Z, Model A) that measures the cost of the error that the model does not explain all of the data. When the predicted variable is continuous, L is usually a squared error measure: (sum of $(Y_{predicted} – Y_{actual})^2$. When predicted variable is discrete (0/1), take # of misclassifications – average error.

IF YOU KNOW THE STATISTICAL PROPERTIES of all the data F(Z) the solution is to select the model A that minimizes $E(L(Z,A) = \int L(Z, A)\, dF(Z)$ where Z covers the entire data space. This is called **the risk of A.**

BUT WE DO NOT KNOW F(Z) and thus do not know the risk. Then use data to calculate the EMPIRICAL RISK of model A: the average error over the observed data (= sum of the errors over Z divided by # observations).

So how do we find out how this model is likely to work in rest of space? Base analysis on two ideas.

Idea 1: Start with a model that does a good job of fitting the observed data. A model with many errors in the observed data is probably not going to fit unobserved data. Maybe we should try a model that **minimizes empirical risk – the loss function over the part of Z we have observed.**

Idea 2: Do not let the model get so specialized to observed data as to over-fit.

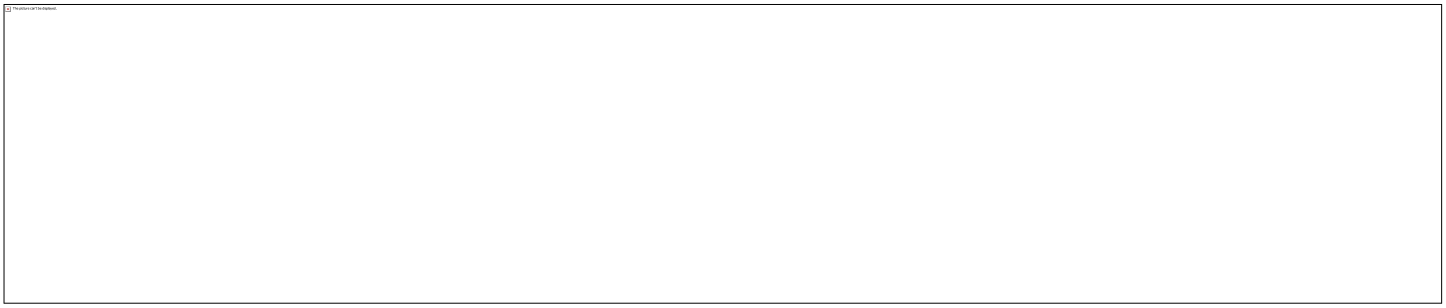
How well does a model that minimizes empirical risk also minimize expected risk? And how much greater is the expected risk to be in the rest of the space than in the observed part of the space?

There are two possible errors in generalizing from the model that fits data to the rest of the world:

Sample error — the difference between the function that fits the data and observed data (this is normal error from any standard statistical analysis) — empirical risk on observed data.

Approximation error — difference between the chosen function and true function — expected risk in world.

Trade-off between these errors. The more complex the function, the better we fit any given data, which reduces the sample error, but the more complex the function, the greater is the risk that it differs from the true function. Need some criterion to say, fit but don't fit too well. Minimizing the empirical risk does not minimize the expected risk if we keep adding more and more terms to the model.



Standard stats estimates a parameter and its standard error from known model. The law of large numbers tells us that as N—> ∞, likely to get good results — asymptotic statistics. Statistical learning data mining estimates a functional relation from a set of possible functions, (which is similar to estimating a parameter, but one level higher. We need a law of large numbers that says the well-chosen estimated function will approach the true function as N—> ∞ as well.

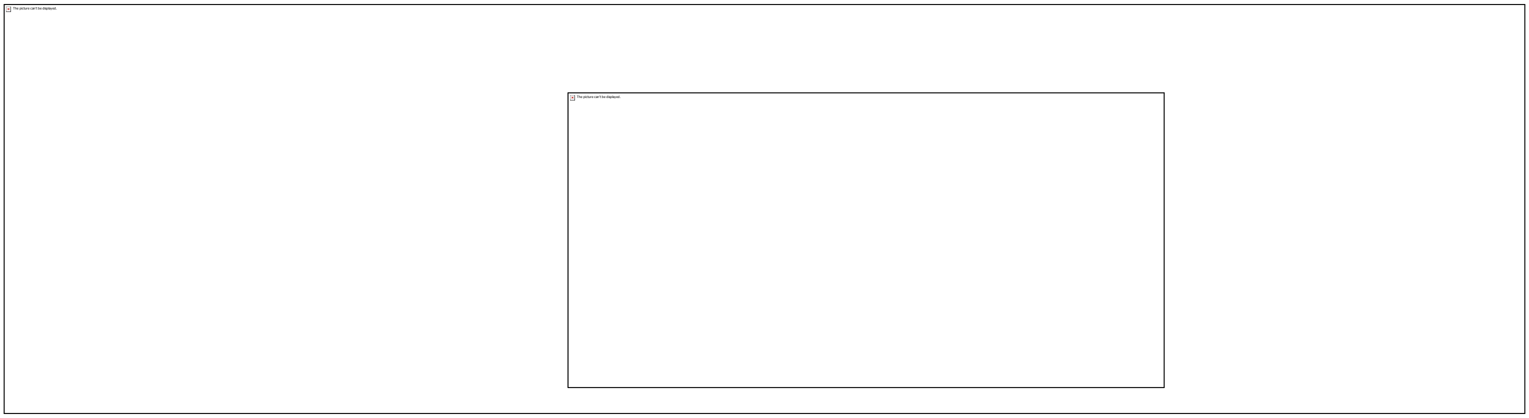
Statistical learning theory identifies **bounds** for the MAXIMUM difference between the expected risk and empirical risk. These bounds are based on small sample characteristics – on measures of the danger of over-fitting given the type of model you have picked (VAPNIK-CHERVONENKIS or VC dimension). This means you can use the data to pick the function that fits it just as you use data to pick parameters of a known function. **The theory says that as the model gets more complicated, you reduce the empirical loss but add to the expected loss**. It gives you a bound on the expected loss — a "cost of complexity" – where more complexity is comparable to more dimensions.

GENERALIZATION ERROR = NOISE + BIAS + VARIANCE

NOISE = intrinsic error that no model will eliminate

BIAS = error due to faulty model -- as model improves, this goes to zero

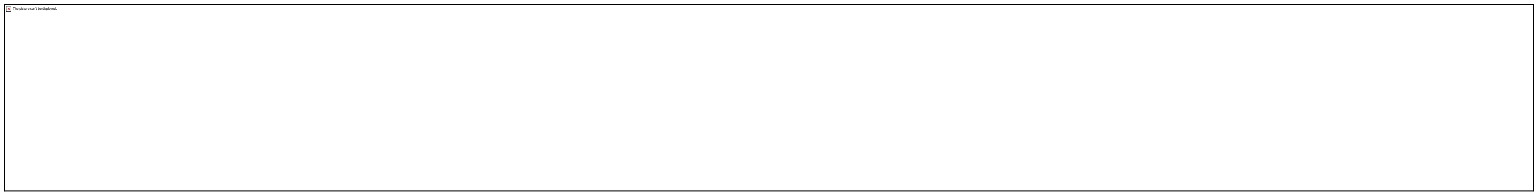
VARIANCE = error due to estimation -- as sample grows this goes to zero


The picture can't be displayed.
The picture can't be displayed.

The above are learning curves. They measure the error rate for the training and test data sets as the complexity/ size/# of training records increases. The training/test division is designed to minimize the generalization error so that when you apply the model to NEW (not training nor test) data it does best job of predicting outcomes.

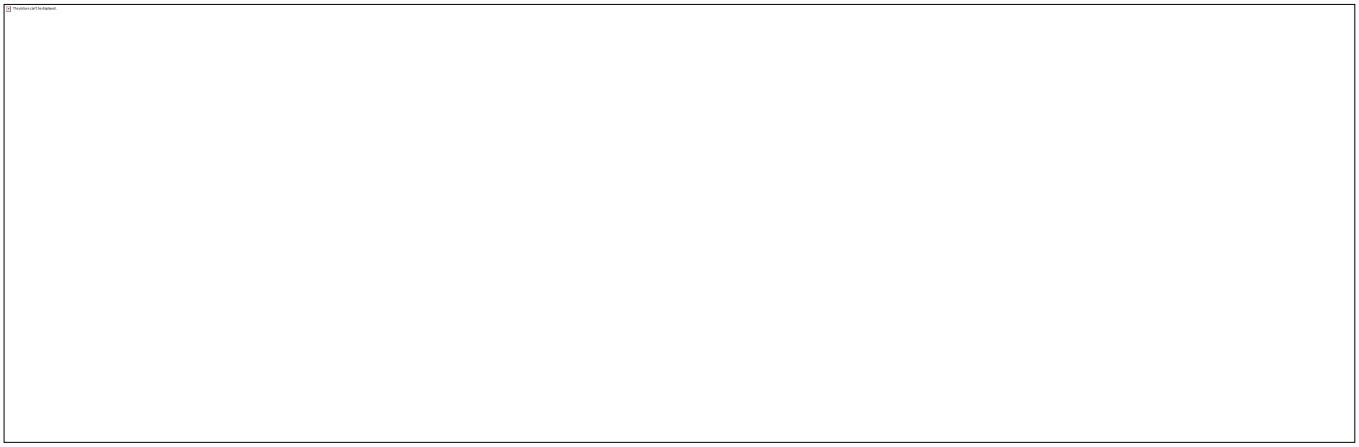
Theorem: Under broad conditions with enough data, the loss from expected risk minimization over the entire space is with high probability *the additive distance* of the loss of the best hypothesis, which depends on the "VC dimension. It measures how complicated the data look relative to the hypothesis. Given VC dimension you can calculate how many observations you need to approximate the true function within any confidence level. **More complicated needs more data.**

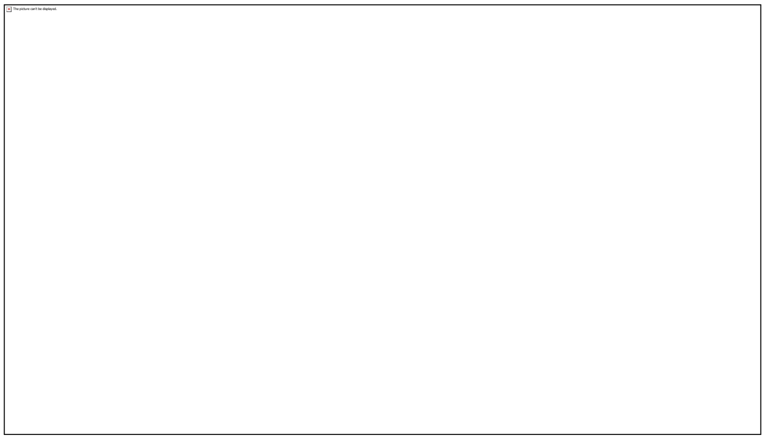
In practice analyst takes the model that minimizes sample error. Then the bound on the error in the TRUE model is the sum of the estimated error in your model and a function that depends on the VC dimension measure of complexity of your model. Put differently, with enough data and a space of possible functions you bound the error over the entire space. The critical assumption is that the unknown space resembles the observed space. The theorem is that with probability 1- η. Test error (on unseen data) depends on error on actual data-mining; h is the VC dimension, the complexity of model; R is size of training set

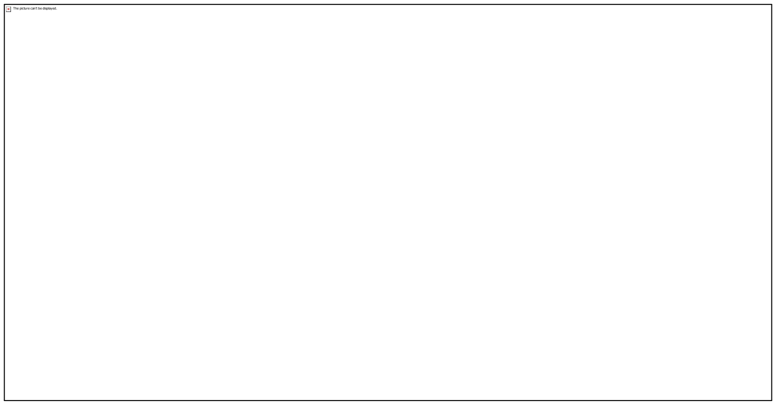


Does this work in practice with real data?

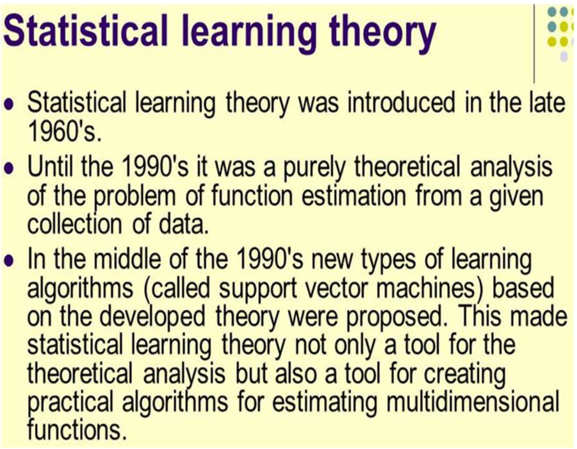
Galindo, Tamayo, Computational Economics, 2000 estimate a tree model for Mexican bank default and find that when sample size increases, test and training error converge. Given that test error asymptotes to Noise and Bias, this allows them to estimate those parameters from the generalization error equation.







The picture can't be displayed.



Statistical learning theory

- Statistical learning theory was introduced in the late 1960's.
- Until the 1990's it was a purely theoretical analysis of the problem of function estimation from a given collection of data.
- In the middle of the 1990's new types of learning algorithms (called support vector machines) based on the developed theory were proposed. This made statistical learning theory not only a tool for the theoretical analysis but also a tool for creating practical algorithms for estimating multidimensional functions.

The picture can't be displayed.