Analyse et Traitement de l’Information

Probabilities and Statistics

High-dimensional Data

Data modelling

Representation spaces

Model and data analysis

Data volume and dimension

Complexity and performance

Representation spaces

- Characteristics / features
- Measure and topology
- Space partition
- Statistical data analysis

Representation spaces

- Data types:
  - Scientific data, industrial, financial, etc
  - Multimedia: text, audio, video

- $C=\{d_1,d_2,\ldots,d_N\}$ a document collection
  - For each document: extract features
  - d_i is represented by a vector of M charactéristics x_i in R^M
  - x_i is the machine view of d_i

- Examples
  - Images: x_i is a histogram: 128 colors: M=128
  - Text: x_i measure occurrence of every word in the vocabulary: M=50’000

Approach

Vectors $\xrightarrow{\text{metric}}$ weight $\xrightarrow{\text{topology}}$ continuity

C, a collection of documents, représented par vectors (points) in a vector space

This is a population in R^M

- To index (organize) this collection, we must understand its structure

→We look for geometric properties of this population

→Notions of distance, neighborhood

→We study the statistical properties of this population

→density, generative law

Representation spaces

$$C = \{x_1, x_2, \ldots, x_N\} \quad x_i \in R^M$$

- We want to induce an order over *C*, some base structure
  - We define a topology over the representation space

→Study neighborhoods

→Define a distance

Norm and distances

$$C = \{x_1, x_2, \ldots, x_N\} \quad x_i \in R^M$$

- Norm
  - $\|x\|$ norm of *x*, vector from R^M
  - $\|x\|^2 = \langle x, x\rangle = x^T x$ if the norm arises from a inner product

    Exple: $\langle x, y\rangle = \sum_{i=1}^{M} x_i . y_i \quad \Rightarrow \quad \|x\| = \sqrt{\sum_{i=1}^{M} x_i^2}$
- Distance (metric)

$$\begin{aligned} & d : C \times C \to R^+ \\ & d(x,x) = 0 \quad \forall x \\ & d(x,y) = d(x,y) \quad \forall x, y \\ & d(x,z) \le d(x,y) + d(y,z) \quad \forall y \end{aligned}$$

- Norm and distance

$$d(x,y) = \|x - y\|$$

Ilustrations: http://www.viz.tamu.edu/faculty/ergun
Wikipedia

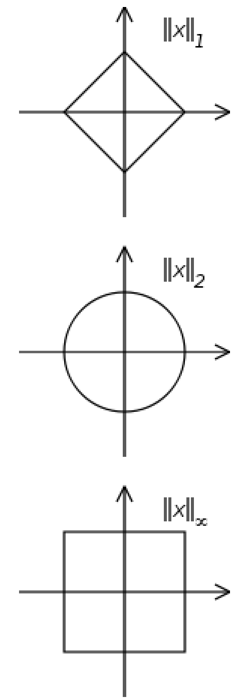
Norms and distances

$$C = \{x_1, x_2, \ldots, x_N\} \qquad x_i \in R^M$$

- Examples of norms (distances)
  - Minkowsky norms (norms L_p)

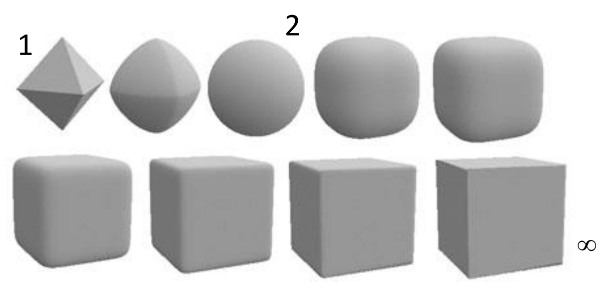
$$\|x\|_p = \left(\sum_{i=1}^{M} x_i^p \right)^{\frac{1}{p}}$$

- p=1 : norm L_1 $\quad \|x\|_1 = \sum_{i=1}^{M} |x_i|$
- p=2: norm L_2 (Euclidean)
- $p = \infty$: norm $\quad L_\infty \quad \|x\|_\infty = \max_i \left(|x_i| \right)$



- Unit Ball: given distance d(.,.)

$$B_d(x) = \{y \text{ t.q } d(x,y) \le 1\} \quad \text{(closed)}$$
$$B_d(x) = \{y \text{ t.q } d(x,y) < 1\} \quad \text{(open)}$$



Norms and distances

- Generalised Euclidean Distance

$$d_G(x,y) = \sqrt{\sum_{i=1}^{M} \frac{1}{w_i}(x_i - y_i)^2}$$

- Mahalanobis Distance

needs to have square root

$$A \in R^{MxM} \quad \text{psd} \quad (x^T A x > 0\,;\, \forall x \neq 0)$$

$$d_A^2(x,y) = (x-y)^T A^{-1}(x-y)$$

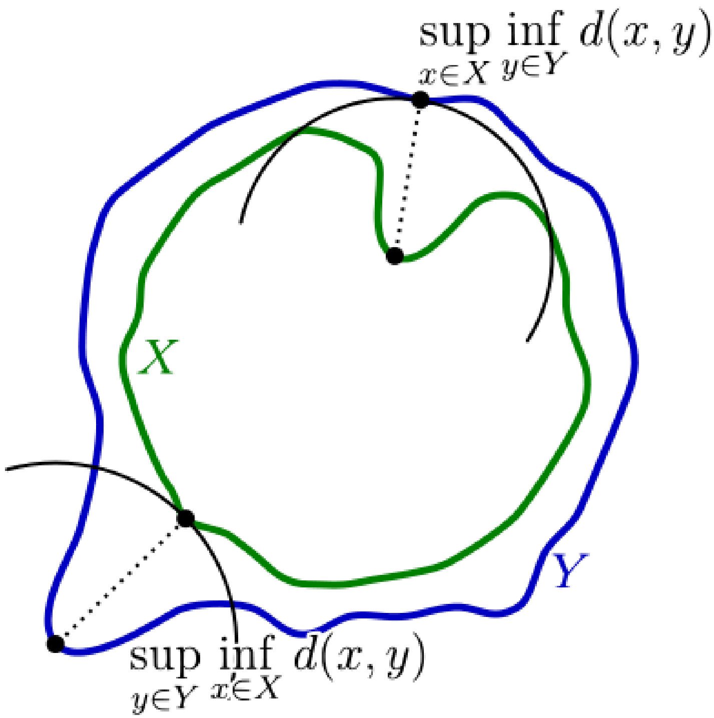
$$\text{si} \quad A = \text{Id} \quad \Rightarrow \quad d_A = d_2$$

$$\text{si} \quad A = diag(w_i) \quad \Rightarrow \quad d_A = d_G$$

Norms and distances

- Hausdorff Distance

  X, Y subsets of *C*



$$d_H(X,Y) = \max(\sup_{y\in Y}\inf_{x\in X} d(x,y), \sup_{x\in X}\inf_{y\in y} d(x,y))$$

Illustration: Wikipedia)

Nearest neighbors

One of the most frequently encountered problems in data analysis

- Given a query vector $q \in R^M$
- We look for its neighborhood *V*

K-NN (nearest neighbor) $k \in N^*$

$$V = \left\{ x_{i_1}, x_{i_2}, \ldots, x_{i_k} \ \text{t.q} \ d(q, x_{i_l}) \le d(q, x_j) \quad \forall j \notin \{i_1, \ldots, i_k\} \right\}$$

x_{i_1} is the closest k-neighbor

x_{i_k} is the farthest k-neighbor

ε-NN ε>0, fixed

$$V = \left\{ x_{i_1}, x_{i_2}, \ldots, x_{i_k} \ \text{t.q} \ d(q, x_{i_l}) \le \varepsilon \quad \forall k \right\}$$

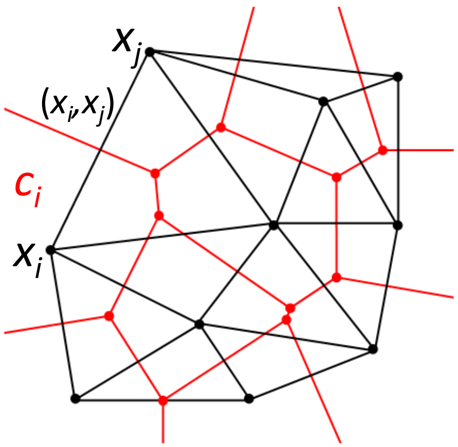
Space partitioning

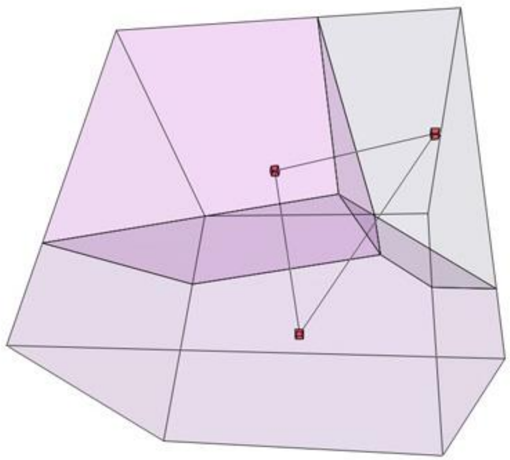
$$C = \{x_1, x_2, \ldots, x_N\} \quad x_i \in R^M$$

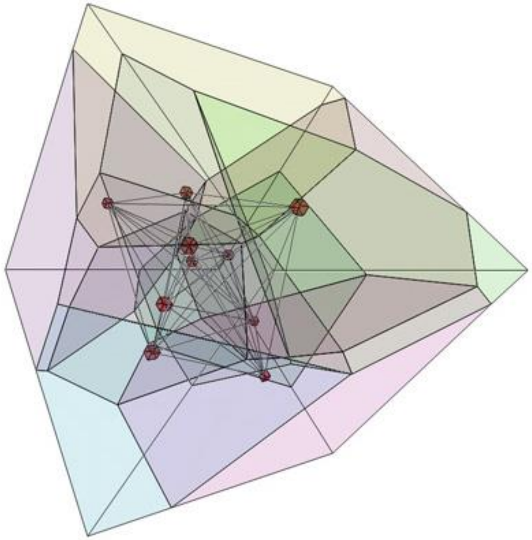
Voronoi Diagram

c_i: Voronoi cell of point x_i

$$c_i = \{y \in R^M \quad \text{t.q.} \quad d(x_i, y) < d(x_j, y) \ \ \forall j \neq i\}$$







Delaunay Graph

$D=(C,E)$: Points x_i are the vertices of D

(x_i,x_j) is an edge iff c_i and c_j share a side

Edges connect neighboring cells

Probability

Given experience *E*, *W* is the set of possible outcomes

A specific outcome of *E* is *w* from *W*

An event $A \subset W$ is a subset of *W* (an assertion)

Examples:

E=« roll a dice » (experience)

W={1,2,3,4,5,6}

w=2

A=« result is even » = {2,4,6}

Probability: function *P* measuring the odds of an event occurring

Probability

P: probability function

$$P(W) = 1$$

$$P(\phi) = 0$$

$$0 \le P(A) \le 1 \quad \forall A$$

$$P(\overline{A}) = 1 - P(A)$$

$$P(A) \le P(B) \quad \text{si } A \subset B$$

$$P(A \cup B) = P(A) + (PB) - P(A \cap B)$$

$$P(\bigcup_i A_i) = \sum_i P(A_i) \text{ if } A_i \text{ are disjoint events}$$

$$\text{if } 0 \le p_i \le 1 \ \forall i \ \text{ and } \sum_{\text{i}} p_i = 1,\ p_i \text{ is probaility of event } w_i \in W$$

$$\text{then} \quad P(A) = \sum_{w_i \in A} p_i$$

$$\text{if } p_i = \text{cste} \ \forall i \quad \text{(equiprobable events)} \ P(A) = \frac{|A|}{|W|}$$

Exple: *P*(« result is odd in a fair dice »)=3/6=0.5

Probability

- Joint probability

$$P(A \cap B) = P(A,B) = P(B \cap A) = P(B,A)$$

- Conditional Probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A,B)}{P(B)}$$

- Independent events
  - if $P(A \cap B) = P(A)P(B)$

    then

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

  Bayes Formula

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Random variable (r.v) – formal definition

A measurable function

$$X : W \to E \quad \text{with } E \text{ measurable (classically } E \subseteq \Re)$$

$\Rightarrow$It is a way to associate labels (which can be *structured* - eg in subsets or intervals) to every event in *W*

$\Rightarrow$We can associate the odd of realization to any set of labels as a measure of their corresponding part

$$P(X = x) = P(X^{-1}(x)) \quad \forall x \in E$$

Exples:

Roll a fair dice: P(*X*="5") = P("face labeled 5") = 1/6

Toss 2 fair coins: P(*X*="Head,Tail")

Random variable (r.v) – informal definition

An event *A* (subset of *W)* may occur or not. This is translated into a logical proposition that may be either True or False

Eg.: *A*=« result is even » → "result is even" = True

A random variable is translated into a logic proposition that has probability *p* to be True and 1-*p* to be False

If the random variable has multiple values, its probability is computed as "one-against-all"

Exple:

Roll a dice: P((*X*="5")=True)=*p*=1/6

→ P((*X*="5")=False)=1-*p*=1-1/6=5/6

Hence:

A random variable is a variable

that can take some values with certain probabilities

Probability law of a real r.v.

The probability law of *X, P_x* is defined by

$$P_X(x) = P(X^{-1}(x)) = P(X = x) \quad \forall x \in E$$

→ It is the histogram of *X*

Discrete r.v: *E* discrete (exple: dice, *E*={1,2,3,4,5,6})

$$P_X(x) = P(X = x) \quad x \in E$$

$$P_X(B) = \sum_{x \in B} P_X(x)$$

$$0 \le P_X(x) \le 1 \quad ; \quad \sum_{x \in S} P_X(x) = 1$$

Continuous r.v: *E* continuous (exple: Temperature, *E*=[-10,50])

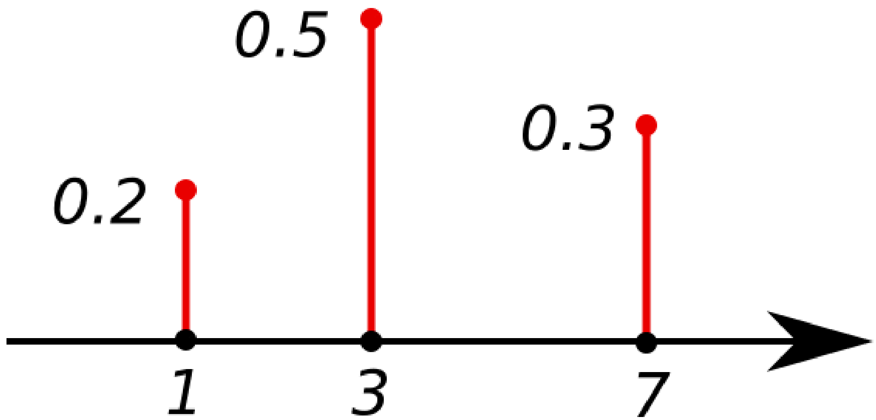
$$P_X(x) = 0 \ ; \quad P_X([x, x+dx[) \ne 0 \quad x \in E$$

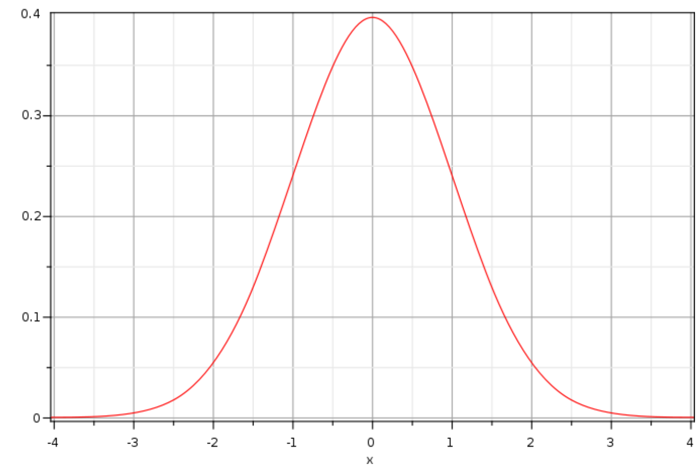
Probability density

$$f(x) = P_X(X \in [x, x+dx[).dx \quad ; \quad f(x) \ge 0 \quad ; \quad \int_R f(x)dx = 1$$

$$P_X(B) = \int_B f(x)dx$$

Laws for r.v





Partition function *F*

- Increasing function

$$F : R \to [0,1]$$

$$x \to P(X \leq x)$$

- Right-continuous function
- $\lim\limits_{x \to -\infty} F(x) = 0 \quad ; \quad \lim\limits_{x \to +\infty} F(x) = 1$

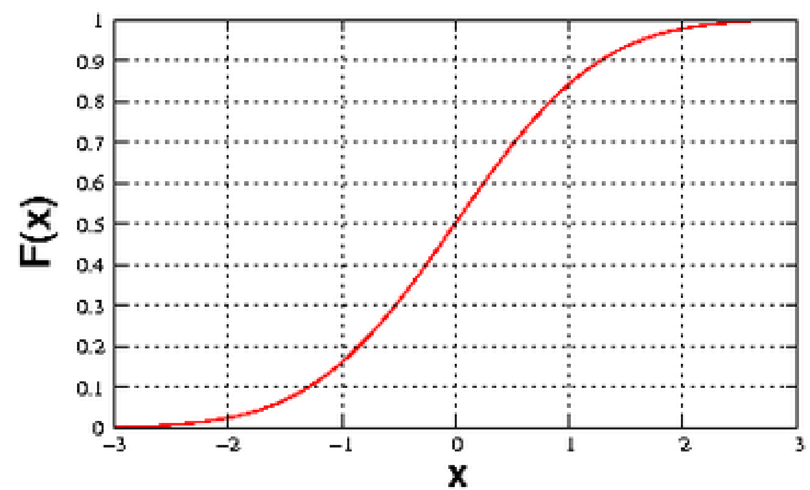


Illustration: Wikipedia

Laws for r.v

Discrete r.v

$$F(x) = \sum_{y \in E; y \leq x} P_X(y)$$

$$P(a \leq X \leq b) = F(b) - F(a)$$

$$P(X > x) = 1 - F(x)$$

Continuous r.v

$$f_X(a) = \frac{\partial F_X(x)}{\partial x} \quad \text{at } a$$

$$F(a) = \int_{-\infty}^{a} f(x)dx$$

$$P(a \leq X \leq b) = F(b) - F(a) = \int_{a}^{b} f(x)dx$$

$$P(X > a) = P(X \geq a) = \int_{a}^{+\infty} f(x)dx = 1 - F(a)$$

Probability

- Expectation

discrete r.v

$$E(X) = \sum_{x \in E} x.P_X(x)$$

$$Y = g(X)$$

$$E(Y) = \sum_{x \in E} g(x)P_X(x)$$

continuous r.v

$$E(X) = \int_R x.f(x)dx$$

$$Y = g(X)$$

$$E(Y) = \int_R g(x).f(x)dx$$

$$E(a) = a \qquad E(aX + bY) = aE(X) + bE(Y)$$

$$E(XY) = E(X)E(Y) \quad \text{if } X \text{ and } Y \text{ are indepedent}$$

- Variance

$$V(X) = \sigma_X^2 = E((X - E(X))^2)$$

$$V(X) \geq 0 \qquad \sigma_X \geq 0$$

$$\text{Ecart type}: \quad \sigma_X = \sqrt{V(X)}$$

$$V(X) = E(X^2) - (E(X))^2$$

$$V(aX + b) = a^2 V(X)$$

$$V(X) = 0 \Rightarrow X = \text{cste}$$

Moments

Moments (non centered) of order *k*

$$m_k = (E(X^k)) \qquad m_1 = E(X)$$

Moments (centered) of order *k*

$$\mu_k = E((X - E(X)^k) \qquad \mu_1 = 0 \quad \mu_2 = Var(X)$$

Symmetric law: $\mu_{2k+1} = 0 \quad \forall k \geq 0$

Asymmetry (skewness)

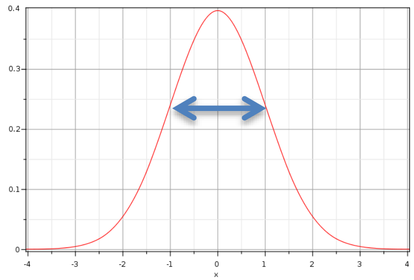
$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

Flatenning (kurtosis)

$$\gamma 2 = \frac{\mu_4}{\sigma^4}$$

Chebichev inequality

- Relationship between the standard deviation and the dispersion around the expectation



$$P(|X-\mu|>n\sigma)\leq\frac{1}{n^2}$$

  - $n=\sqrt{2}$: at least half of the values are in $[\mu-\sqrt{2}\sigma,\mu+\sqrt{2}\sigma]$
  - Gaussian *N*(0,1) : $P(|X|<3)\cong 0.9973$

- Normalized and centered variable:

$$X^*=\frac{X-E(X)}{\sigma_X}\quad;\quad E(X^*)=0\quad;\quad\sigma_{X^*}=1$$

Illustration: Wikipedia

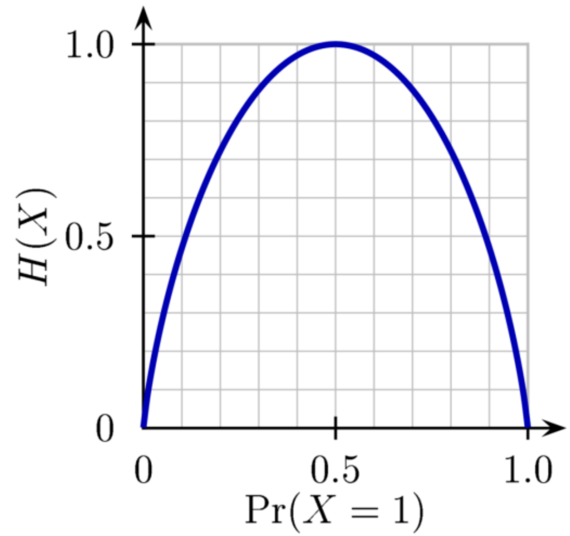
Entropy

- The Entropy of a discrete r.v X is

$$H(X) = E(-\ln(P(X)) = E(I(X)) \quad \text{where} \quad I(X) = -\ln(P(X))$$

$I(X)$ estimate the information content in X (in bits)

$$H(X) = \sum_{i=1}^{n} P(x_i).I(x_i) = -\sum_{i=1}^{n} P(x_i).\ln(P(x_i))$$



Entropy vs bias of a coin

Illustration: Wikipedia

Pairs of r.v

$$\text{Discrete}: \quad P(\{X=x\}\cap\{Y=y\})_{x\in E, y\in F}$$

$$\text{Continuous}: \quad f(x,y)dxdy = P(\{X\in[x,x+dx[\}\cap\{Y\in[y,y+dy[\})$$

Marginal law

$$\text{Discrete}: \quad P(\{X=x\}) = \sum_{y\in F} P(X=x, Y=y)$$

$$\text{Continuous}: \quad f(x) = \int_{-\infty}^{+\infty} f(x,y)dy$$

Pairs of r.v

- Conditional laws

$$P(Y=y \mid X=x)=\frac{P(X=x,Y=y)}{P(X=x)} \quad \forall y \in F$$

$$f(y \mid x)=\frac{f(x,y)}{f(x)} \quad \forall y \in R$$

- Covariance

$$\operatorname{cov}(X,Y)=E((X-E(X))(Y-E(Y)))$$

$$\Rightarrow \operatorname{cov}(X,Y)=E(XY)-E(X)E(Y)$$

$$\Sigma(X,Y)=\begin{pmatrix} V(X) & \operatorname{cov}(X,Y) \\ \operatorname{cov}(X,Y) & V(Y) \end{pmatrix}$$

- Correlation

$$\rho(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$$\Rightarrow \rho(X,Y)=\operatorname{cov}(X^*,Y^*)$$

$$-1 \le \rho(X,Y) \le 1$$

$$|\rho(X,Y)|=1 \quad \Rightarrow \quad Y=aX+b$$

$$\rho(X,Y)=0 \quad \Rightarrow \quad \operatorname{cov}(X,Y)=0$$

$$\Rightarrow \quad V(XY)=V(X)V(Y) \qquad E(XY)=E(X)E(Y)$$

Usual discrete laws

$$C_n^x = \frac{n!}{x!(n-x)!}$$

- Bernoulli:
  - Draw from an urn containing a proportion of *p* white balls and *q*=1-*p* red balls. *X*=« number of red balls »

$$x \in \{0,1\} \quad Ber(p): P(X=x) = p^x q^{1-x} \quad x \in \{0,1\} \quad E(X) = p \quad V(X) = pq$$

- Uniform
  - fair dice with *n* faces

$$U(n): P(X=x) = \frac{1}{n} \quad x \in \{1,\ldots,n\} \quad E(X) = \frac{n+1}{2} \quad V(X) = \frac{n^2-1}{2}$$

- Binomial
  - *n* draws with return from a Bernoulli urn

$$Bi(n): P(X=x) = C_n^x p^x q^{n-x} \quad x \in \{0,1\} \quad E(X) = np \quad V(X) = nqp$$

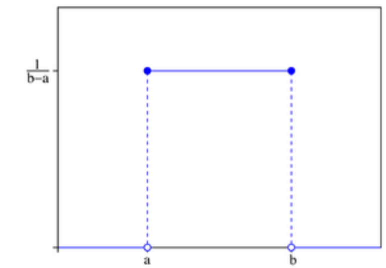
- Poisson
  - Number of people at a bus stop after time λ

$$P(\lambda): P(X=x) = e^{-\lambda}\frac{\lambda^2}{x!} \quad x \in N \quad E(X) = V(X) = \lambda$$

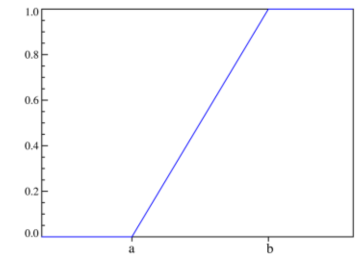
Usual continuous laws

- Uniform

$$U([a,b]):\ f(x)=\frac{1}{b-a}1_{a\le x\le b}\qquad E(X)=\frac{a+b}{2}\qquad V(X)=\frac{(b-a)^2}{12}$$

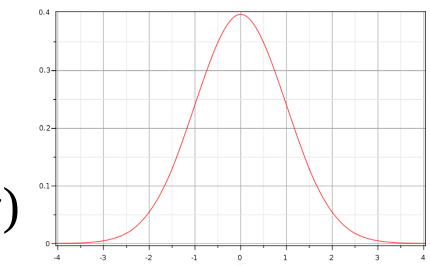


$$F(X)=\begin{cases}0 & si\quad x<a\\ \dfrac{x-a}{b-a} & si\quad a\le x\le b\\ 1 & si\quad x>b\end{cases}$$



- Normal

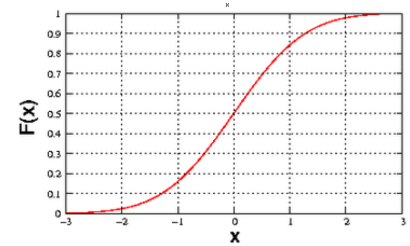
$$N(\mu,\sigma):\ f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\qquad E(X)=\mu\quad V(X)=\sigma^2$$



$$\mu=0\,;\sigma=1\Rightarrow F(0)=\frac{1}{2}\qquad F(x)<\frac{1}{2}\Rightarrow x<0\qquad F(-x)=1-F(x)$$

$$P(|X|<x)=2F(x)-1$$

$$P(|X|<3)\cong 0.9973$$

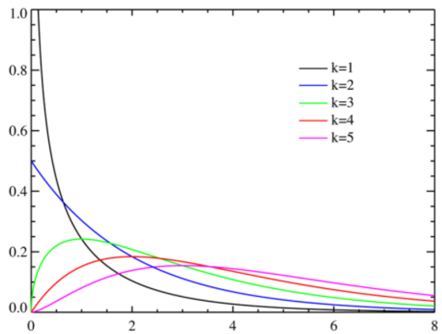


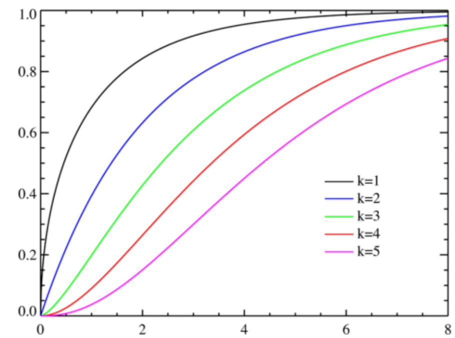
$$N_D(\mu,\Sigma):\ f(x)=\frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

Illustration: Wikipedia

Usual continuous laws

- Chi-2 $$X \propto Z = \sum_{i=1}^{k} X_i^2 \quad X_i \propto N(0,1) \quad ; \quad E(X) = k \quad V(X) = 2k$$

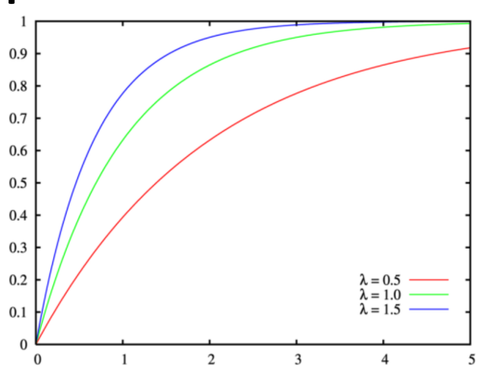


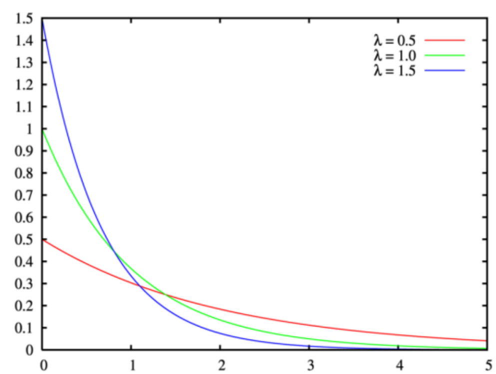


$$F(x) = \frac{\gamma(k/2, x/2)}{\Gamma(k/2)}$$

$$\chi^2(k) : f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} 1_{x \geq 0} \quad k \in N^*$$

- Exponential





$$E(X) = \frac{1}{\lambda}$$

$$V(X) = \frac{1}{\lambda^2}$$

$$Exp(\lambda) : f(x) = \lambda e^{-\lambda x} \quad \lambda > 0$$

$$F(x) = 1 - e^{-\lambda x}$$

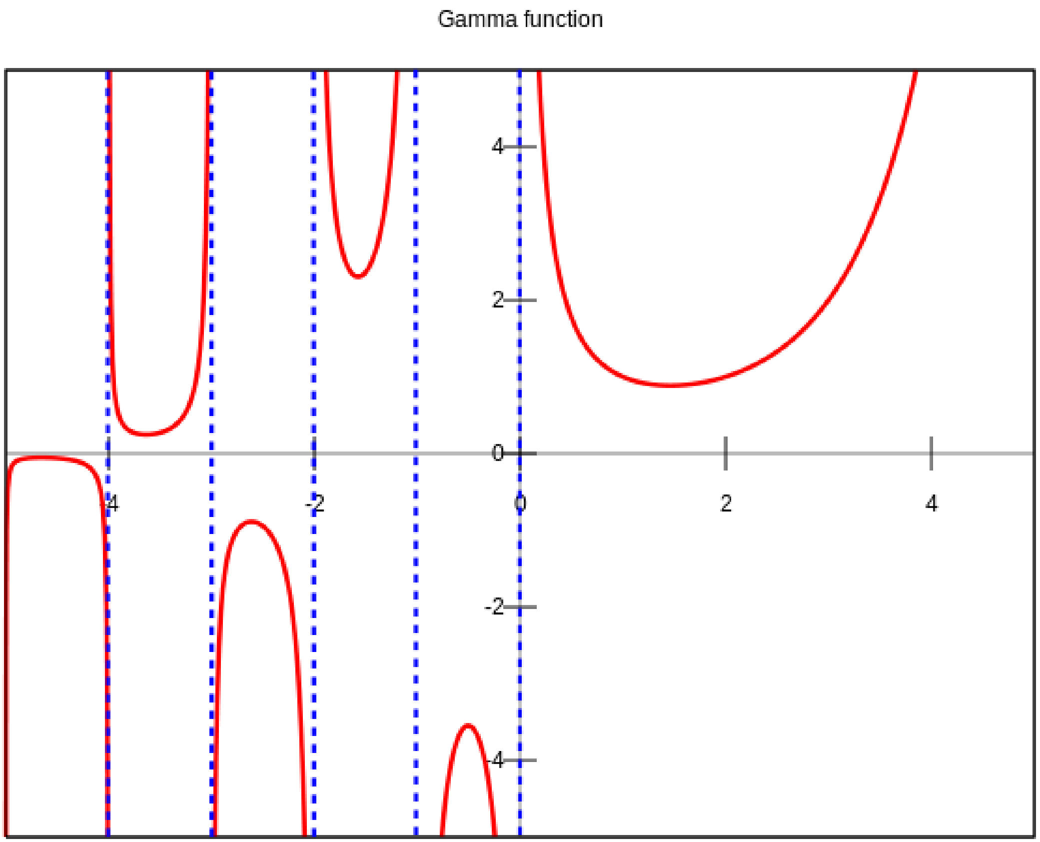
Illustration: Wikipedia

Gamma function (Γ)

- Extension of factorial numbers

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$$

$$\Gamma(n) = (n-1)!$$



To be read in tables or as a function in libraries

Y = gamma(X) in Matlab

Illustration: Wikipedia

Sampling a distribution

- Inversion theorem: F partition function

$$F^{-1}(y) = \inf\{x \in R \quad s.t. \quad F(x) = y\} \qquad U \text{ uniform over } [0,1]$$

(a) The partition function of $X \sim F^{-1}(U)$ is F

(b) If F is continuous over R and X has partition function F,
$U=F(X)$ is uniformly distributed over [0,1]

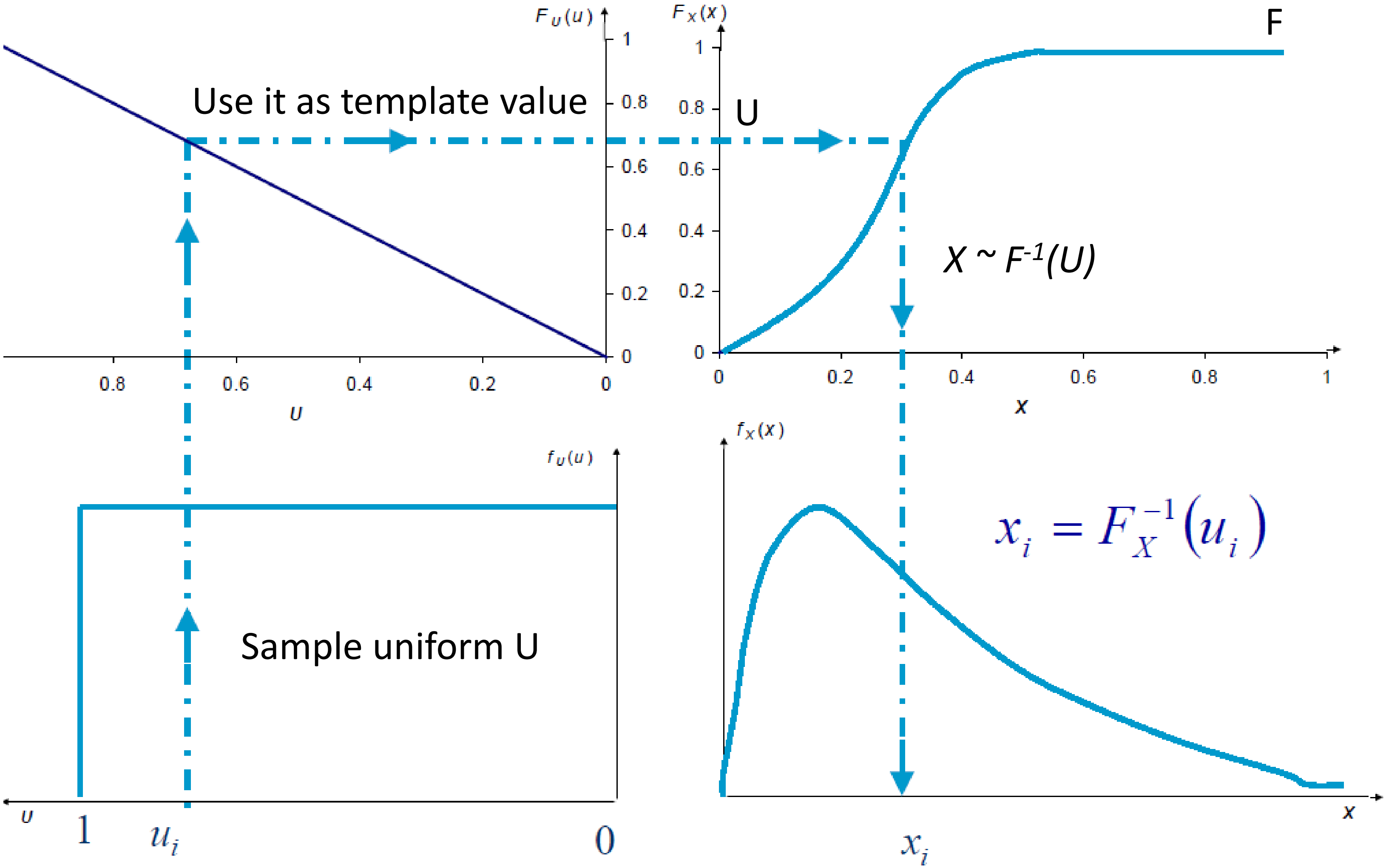
- Sampling
  - Sample n instance $u_1 \ldots u_n$ from a [0,1] uniform distribution
  - Compute $x_i = F^{-1}(u_i)$: these are n samples of X of law f

- Discrete case

$$(p_i = P(X = x_i))_{i=1\ldots n} \quad ; \quad s_k = P(X \leq x_k) = \sum_{i=1}^{k} p_i \quad ; \quad F(u) = \sum_k s_{k-1} 1_{x_{k-1} \leq u \leq x_k}$$

$$\text{if } u_1 \ldots u_n \text{ uniform samples over } [0,1] \quad x_k^* = F^{-1}(u_i) = \sum_{k=1}^{n} x_k 1_{s_{k-1} \leq u_i \leq s_k}$$

Sampling a distribution



Sampling a distribution

In practice F^{-1} is hard to compute

Distribution			
Densité	$F(x)$	$X = F^{-1}$	Forme simplifiée
Exponentielle (λ)			
$\lambda e^{-lx}, x \geq 0$	$1 - e^{-lx}$	$-\frac{1}{\lambda}\ln(1-U)$	$-\frac{1}{\lambda}\ln(U)$
Cauchy[4] (σ)			
$\frac{\sigma}{\pi(x^2+\sigma^2)}$	$\frac{1}{\pi} + \frac{1}{\pi} Arc\tan\left(\frac{x}{\sigma}\right)$	$\sigma\tan\left(\pi\left(U - \frac{1}{2}\right)\right)$	$\sigma\tan(\pi U)$
Pareto $(a, b), b > 0$			
$\frac{ab^a}{x^{a+1}}, w \geq b > 0$	$1 - \left(\frac{b}{x}\right)^a$	$\frac{b}{(1-U)^{1/a}}$	$\frac{b}{(U)^{1/a}}$

For sampling the Gaussian distribution, we can use the Central Limit Theorem (later)

Weak Law of Large Numbers (WLLN)

X a r.v such that $E(X)=\mu$ and $V(X)=\sigma^2$

$(X_1,\ldots,X_n)$ populations, $(x_1,\ldots,x_m)$ samples

then

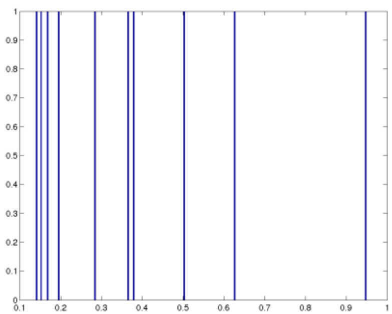
$$\forall n \in N^* \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is an « estimator » of $E(X)$

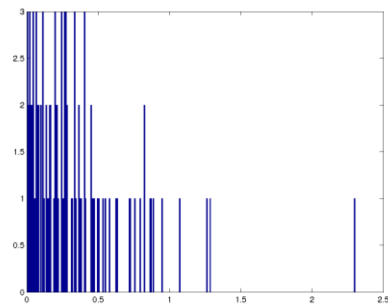
$$\lim_{n\to\infty} P(\left|\overline{X}-\mu\right| \geq \varepsilon) = 0 \; \forall \varepsilon > 0 \qquad E(\overline{X}) = \mu \qquad V(\overline{X}) = \frac{\sigma^2}{n}$$

Sampling an exponential distribution

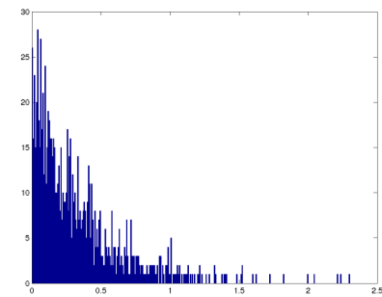
- m samples of uniform law $U([0,1])$
- $X = -\frac{1}{\lambda}\ln(U) \qquad \lambda = 3$



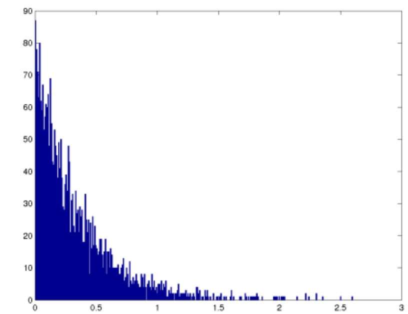
m=10



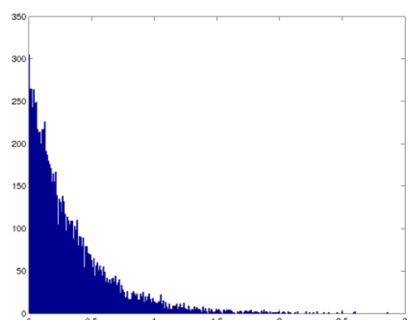
m=100



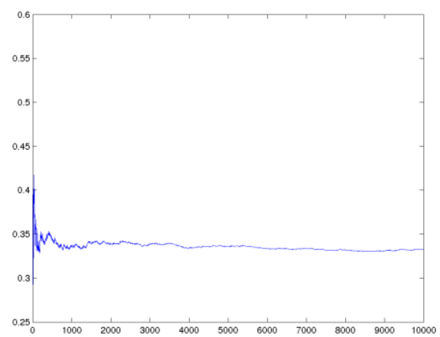
m=1’000



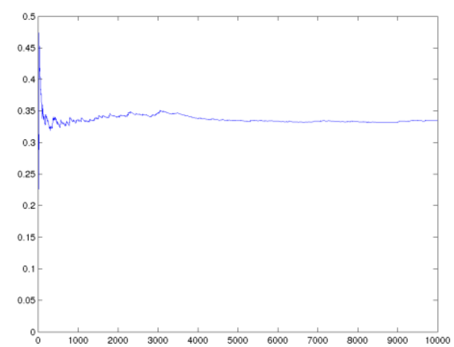
m=3’000



m=10’000



Expectation



Standard deviation

Central Limit Theorem (CLT)

X a r.v such that $E(X)=\mu$ and $V(X)=\sigma^2$

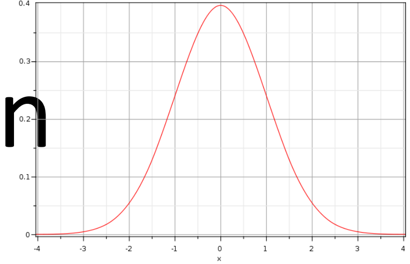
$X_1,\ldots,X_n$ r.v.s following the same law as X

$$\forall n \in N^* \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad Z_n = \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu)$$

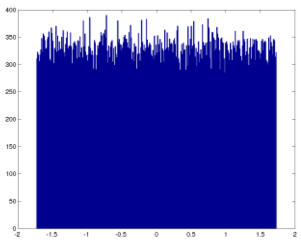
then the law of Z_n converges to the Normal distribution $N(0,1)$

$$\Rightarrow \lim_{n\to\infty} P(a < Z_n < b) = \int_a^b \frac{1}{\sigma\sqrt{2}} e^{-\frac{x^2}{2}} dx$$

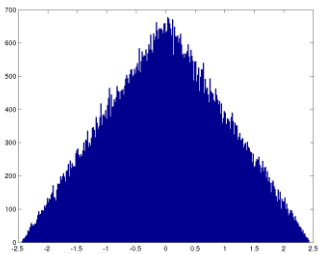
Sampling the Normal distribution



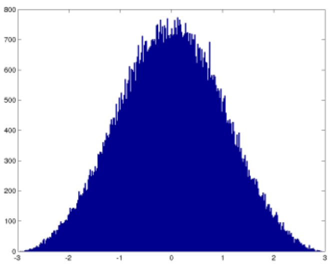
- *m* samples of *n* uniform laws U([-0.5,0.5])
- Average the *n* laws: *m* samples of $\overline{X}$



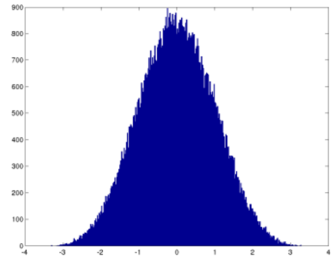
n=1



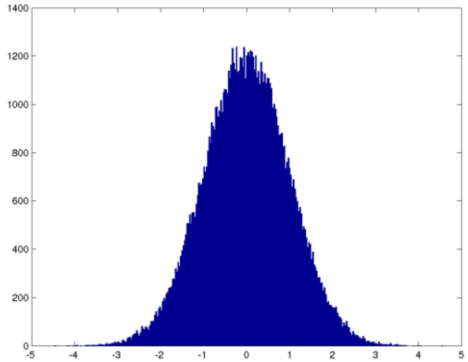
n=2



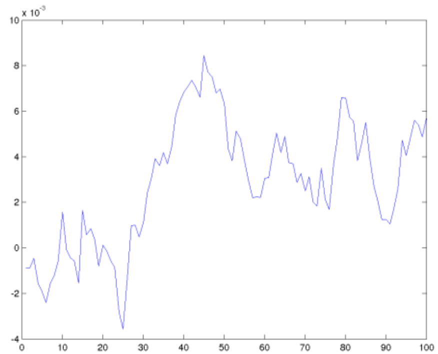
n=3



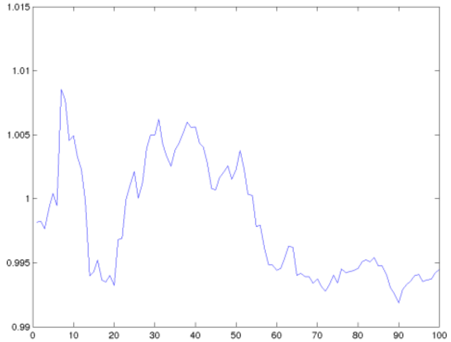
n=4



n=100



Moyenne de Z_n



Ecart-type de Z_n

m=100'000

Interpretation

- X r.v of which μ is to be estimated
  - Exple: « Diameter »
- X_i population
  - Exple: « meter *i* measuring apples »
- x_i : results of the measure
  - Exple: « measured diameters »

→ $\overline{X}$ (the average of meters) converges towards X (by the WLLN)

- The CLT tells us that the error on μ (Z_n) follows a Normal law $N(0,1)$

  Z_n is a r.v representing the average error made by $\overline{X}$

$$\overline{\mu} = \mu \pm Z_n$$

Density estimation

$\{x_1, x_2, \ldots, x_n\}$ *n* samples of a unknown probability density function *f*. We want to estimate the structure of *f*

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

K Kernel function (Noyau)

- *K*(*x*): Symmetric Kernel, integrating to 1
- h>0 : Bandwith (*Bande passante*)

Exple:

- Gaussian Kernel

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

- Pb: How to select *h?*

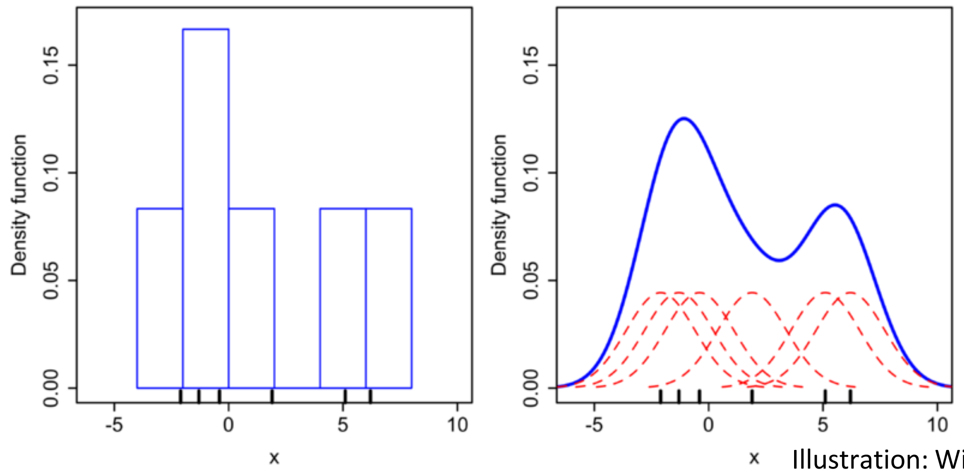


Illustration: Wikipedia

High dimensionality

$$C = \{x_1, x_2, \ldots, x_N\} \quad x_i \in R^M$$

- *M* is the data dimension
  - Measurements, features, …
- *C* is a sample of a *M-dimensional space*

→We wish to study what happens when M increases
  - Influence on geometric notion (distances, *k*-NN)
  - Influence on statistical notions

→« Curse of dimensionality »

→Richard Ernest Bellman (1961). *Adaptive control processes: a guided tour*. Princeton University Press.

→« malédiction de la dimensionnalité »

→ but also "blessing of dimensionality"

High dimensions

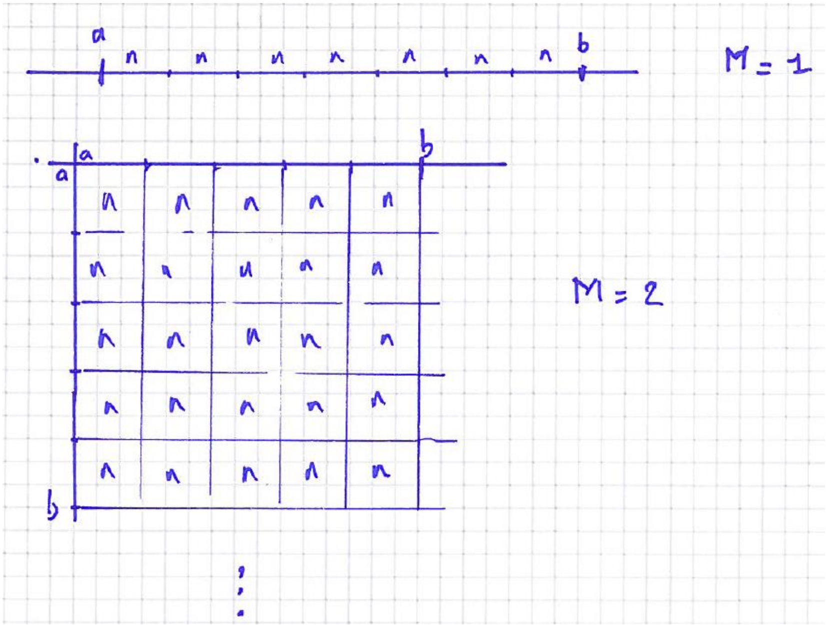
Imagine a population following a distribution in interval $[a,b]^M$

Each dimension is quantized into *k bins*

To estimate the prob law we want *n* samples in each *bin* in average

- $M=1$: $N \sim k.n$
- $M=2$: $N \sim n.k^2$

  …
- M: $N \sim n.k^M$



Exple:

$k=10$, $n=10$, $M=6$ => N ~ 10’000’000 samples

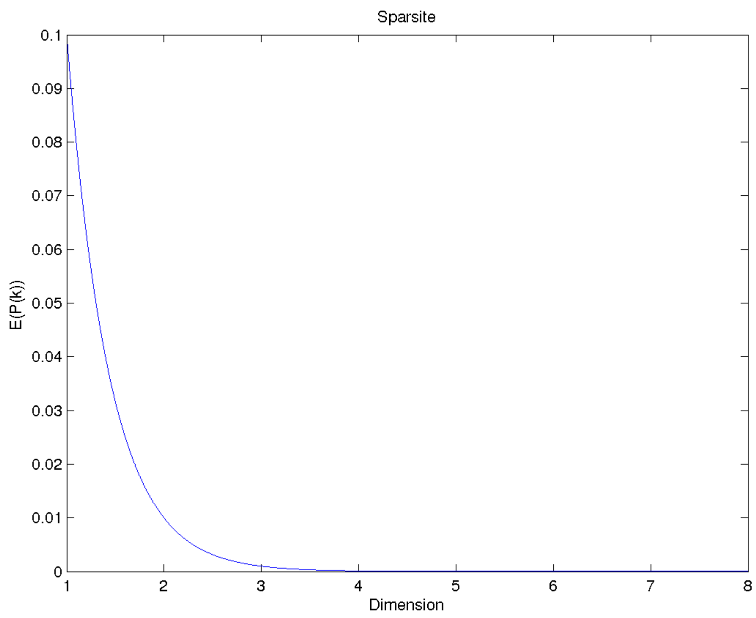
Data sparsity

If we fixe the number *N* of samples:

- *M*=1: $n \sim \frac{N}{k}$ $E(P(x_i \in \mathrm{bin}_k)) \sim \frac{N}{k}$

…

- *M* $n \sim \frac{N}{k^M}$ $E(P(x_i \in \mathrm{bin}_k)) \sim \frac{N}{k^M}$



Structure of a sample

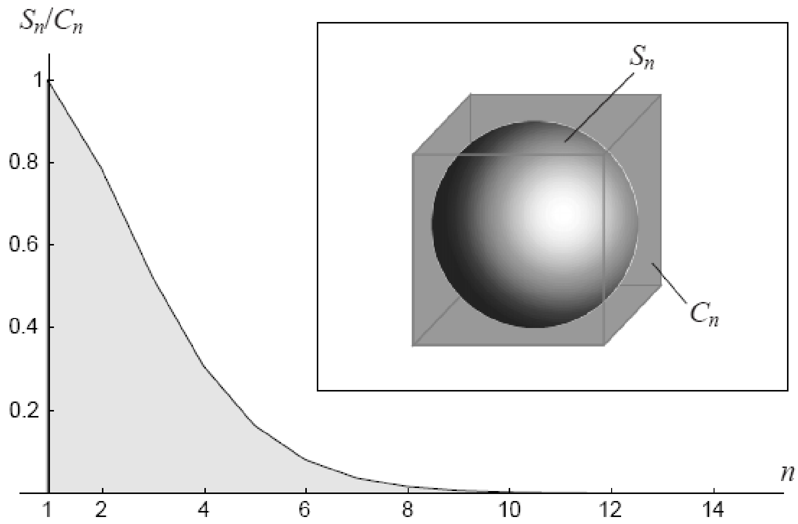
Given S the hypersphere centred at 0 of radius r included in the cube $[-r,+r]^M$

We draw N samples uniformly distributed in $U([-r,+r]^M)$

We can compute the proportion of these samples falling into S

$$V_S(M) = \frac{2r^M \pi^{M/2}}{M\Gamma(M/2)} \qquad V_C(M) = (2r)^M$$

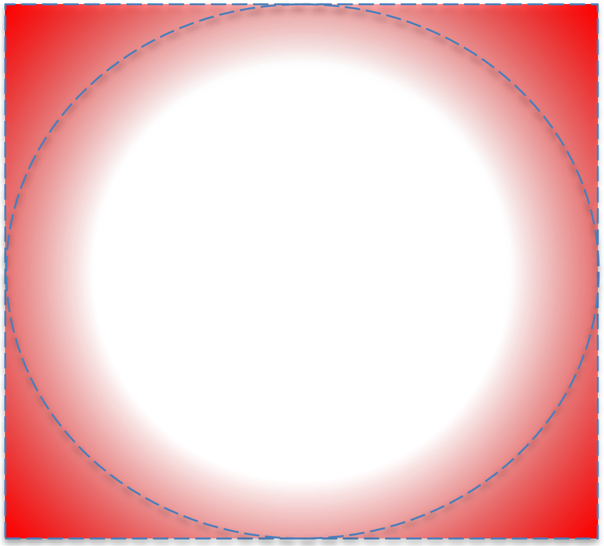
$$\text{ratio} = \frac{V_S(M)}{V_C(M)} = \frac{\pi^{M/2}}{M\,2^{M-1}\Gamma(M/2)} \xrightarrow{M\to\infty} 0$$



Interpretation

In high dimensions (M large – in fact > 10)

- The (relative) volume of the sphere goes to 0
- All samples are in the « corners » of the cube
- All samples « go away from the center »



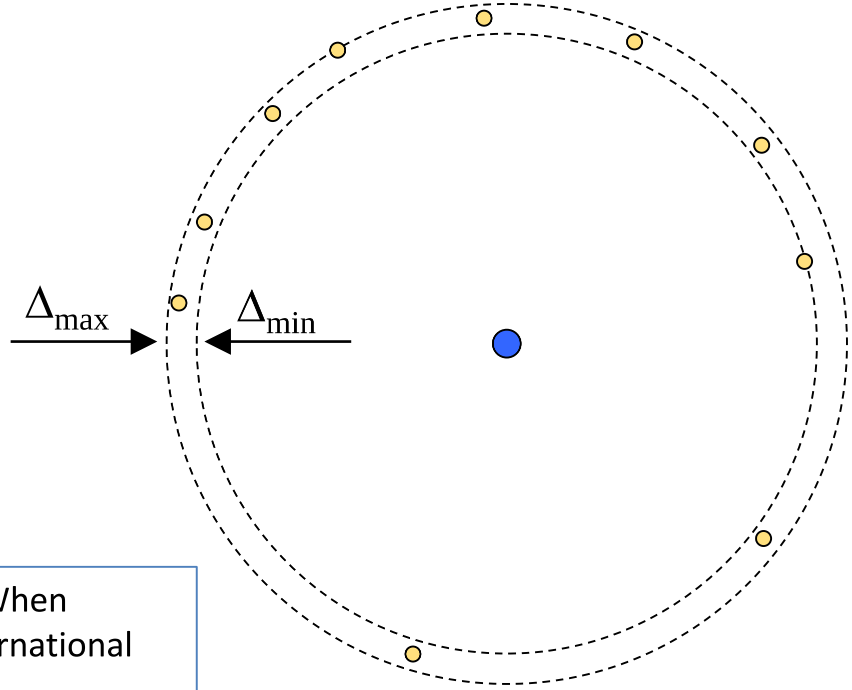
High-dimensional *k*-NN

Δ_{max} and Δ_{min} distances of farthest and closest k-neighbors, respectively. One can prove:

Thm [Beyer et al, 1999]

$$\text{if} \quad \lim_{M\to\infty} V\left(\frac{\|X_M\|_k}{E(\|X_M\|_k)}\right) = 0 \qquad \text{then} \quad \lim_{M\to\infty} \mathrm{P}\left(\frac{(\Delta_{max} - \Delta_{min})}{\Delta_{min}} < \varepsilon\right) = 1 \quad \forall \varepsilon > 0$$

$$\frac{(\Delta_{max} - \Delta_{min})}{\Delta_{min}} \xrightarrow[P]{M\to\infty} 0$$



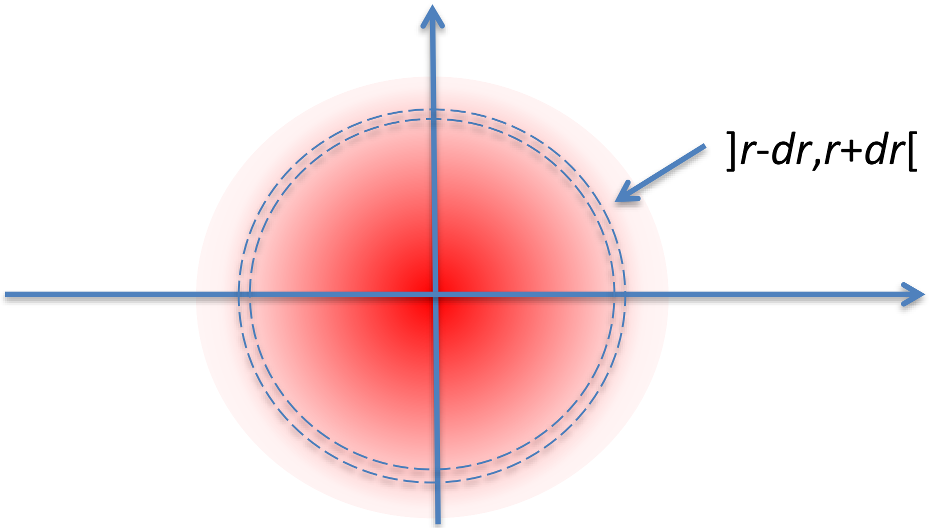
$\Rightarrow$ Neighboring structures are no longer relevant in high dimensions

- $\Rightarrow$ ε-NN: all points are neighbors
- $\Rightarrow$ *k*-NN: random choice

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is“nearest neighbor” meaningful? In Proceedings of the 7th International Conference on Database Theory, pages 217–235

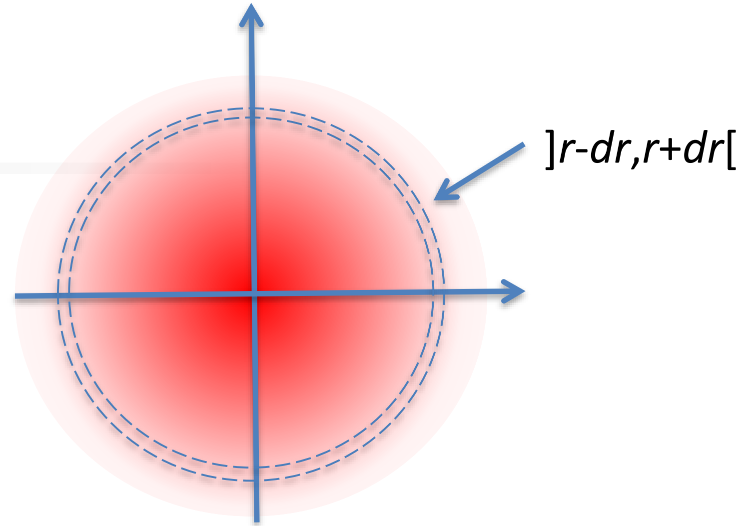
Gaussian distribution

Given a Gaussian distribution in M dimensions $N(\mu_M,\Sigma_M)$, we measure the density in the surface at radius r



WLOG we use the centered scaled distribution $N(0,\mathrm{Id}_M)$

Gaussian distribution



$$X = (X_1,\ldots,X_M) \qquad X_i \sim N(0,1)$$

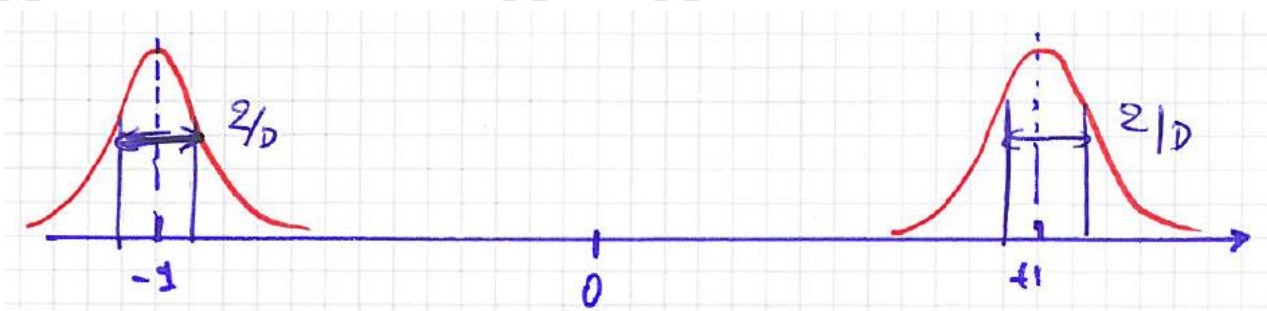
estimation of $\quad P(X = (r,\ldots,r)^T)$

$$\|X\|_2^2 = \sum_{i=1}^{M} X_i^2 = M.r^2$$

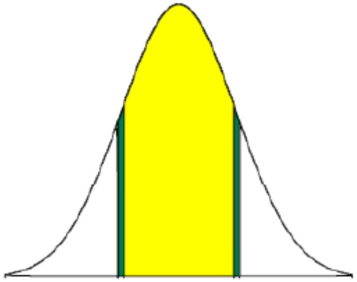
$$\Rightarrow r^2 = \frac{1}{M}\sum_{i=1}^{M} X_i^2 \sim \frac{1}{M}\chi^2$$

$$E(aX+b) = aE(X)+b$$
$$V(aX+b) = a2V(X)$$
$$E(\chi^2) = k \quad ; \quad V(\chi^2) = 2k$$

$$\Rightarrow E(r^2) = \frac{1}{M}M = 1 \qquad V(r^2) = \frac{2M}{M^2} = \frac{2}{M}$$



Gaussian distribution



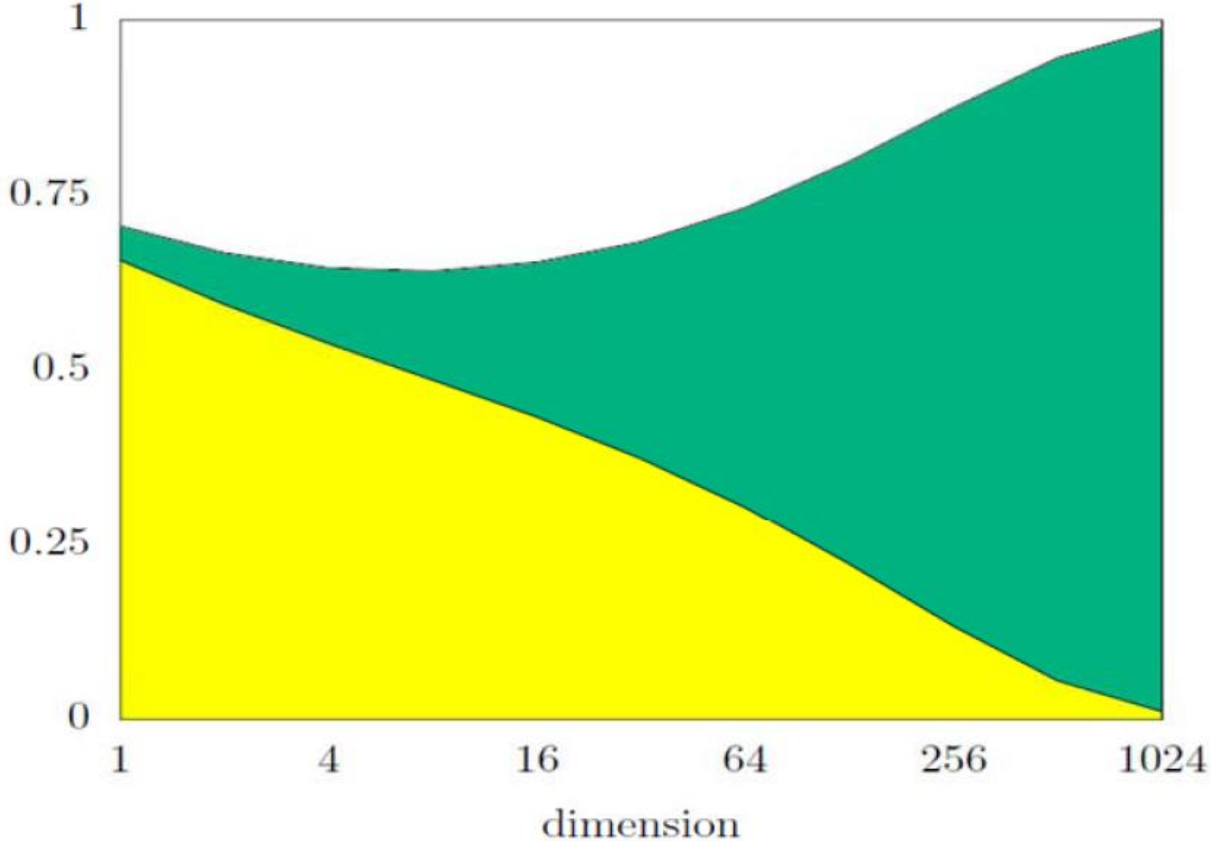
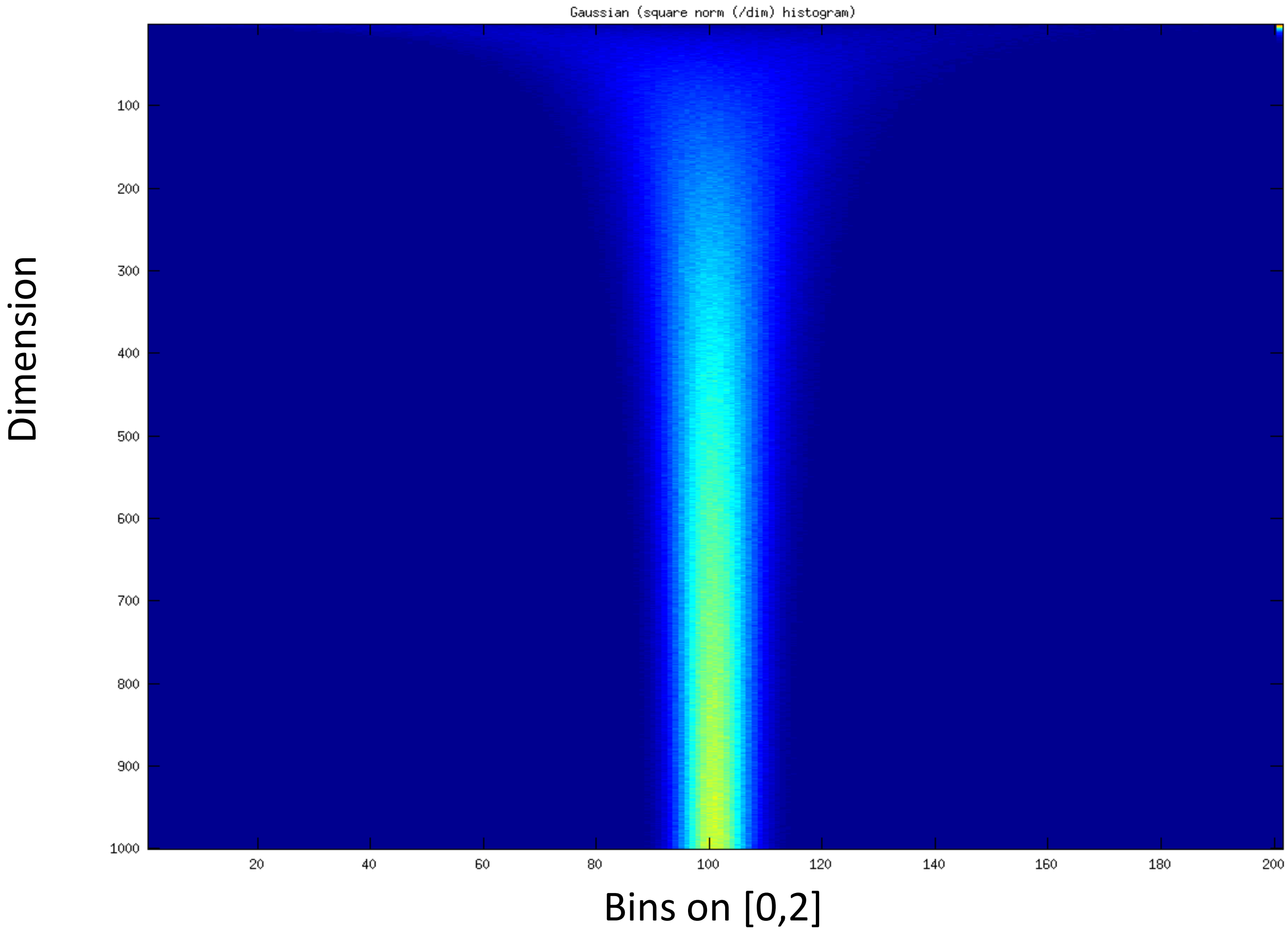
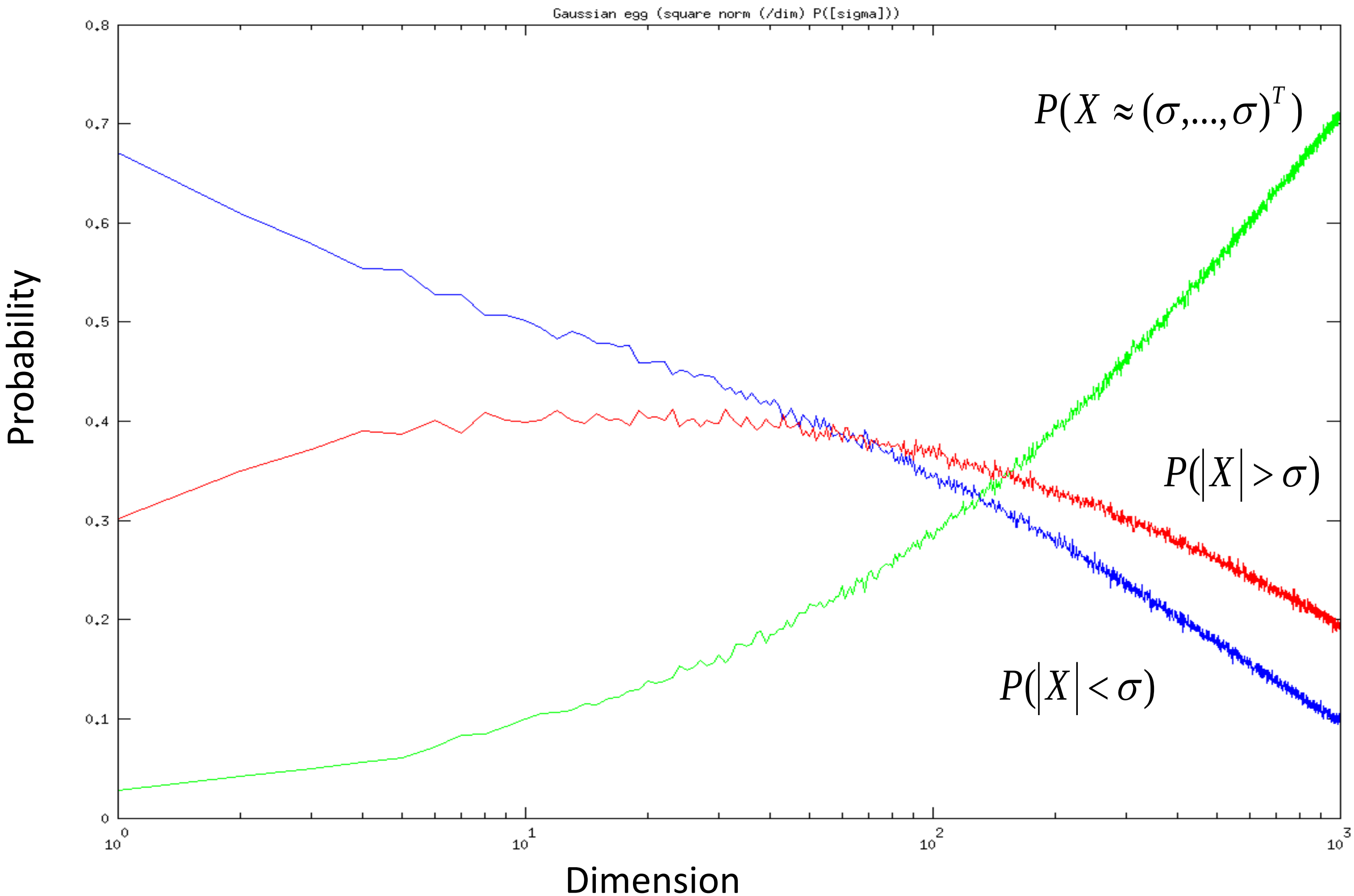


Figure 1.5: Plot of probability mass versus dimension. Plot shows the volume of density inside 0.95 of a standard deviation (yellow), between 0.95 and 1.05 standard deviations (green), between 1.05 and 2 standard deviations (white)

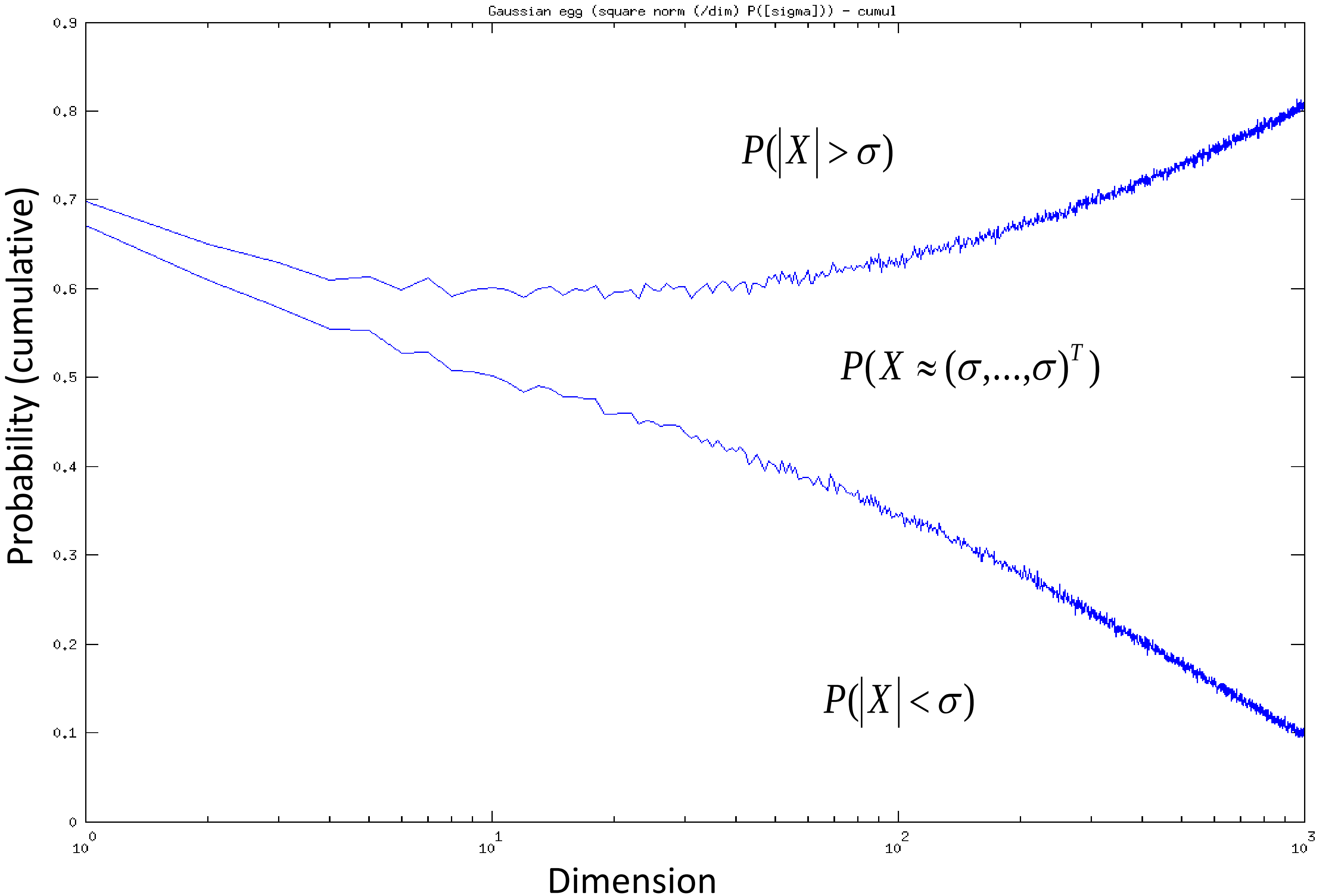
Empirical evidence (10’000 samples)



Empirical evidence (10’000 samples)



Empirical evidence (10'000 samples)



Hubs

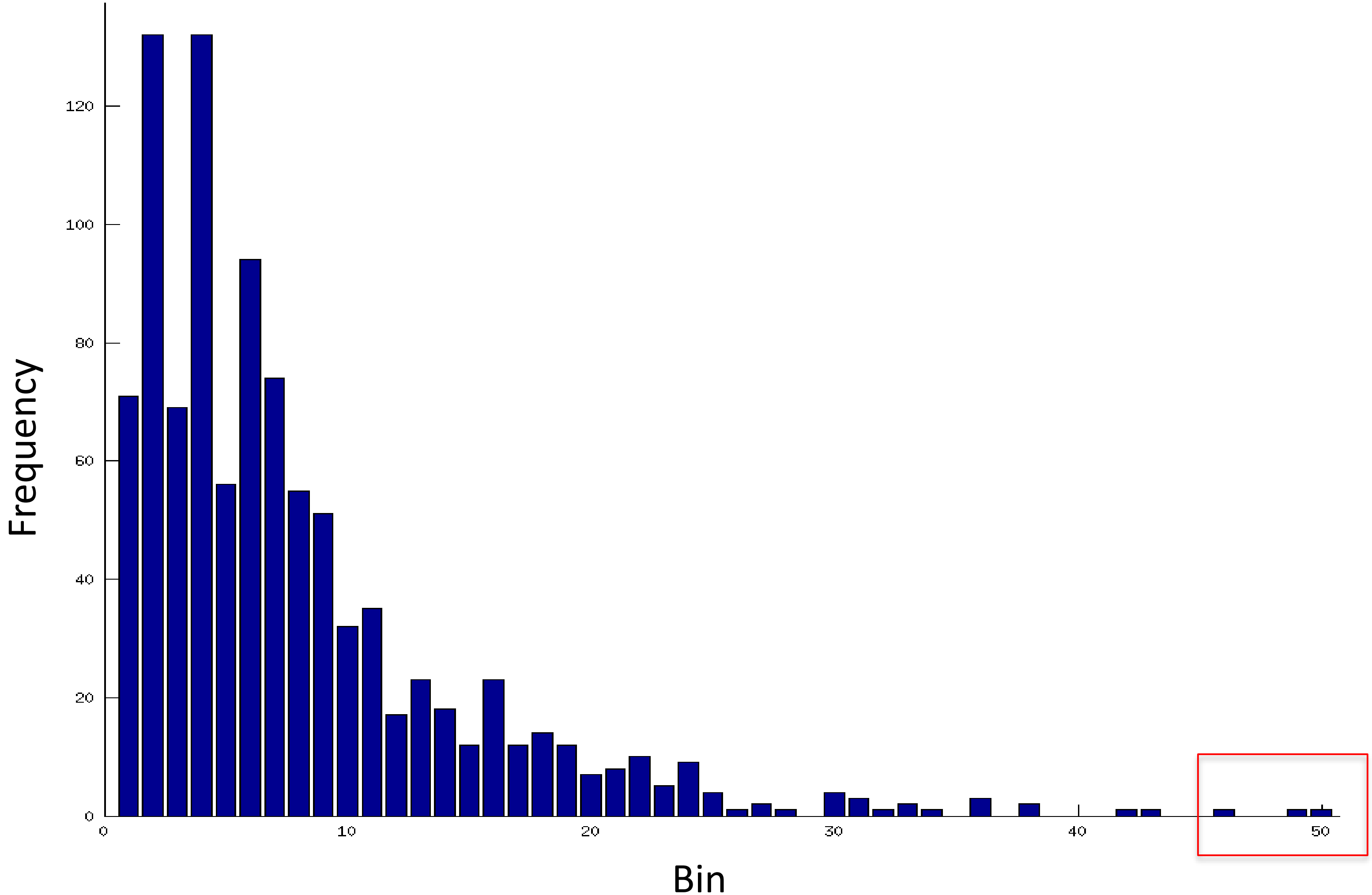
- We want to characterise the number of times a sample appears in the k-NN of another sample:

$$P_{ik}(x) = \begin{cases} 1 \text{ if } \ x \in \mathrm{NN}_k(x_i) \\ 0 \text{ otherwise} \end{cases}$$

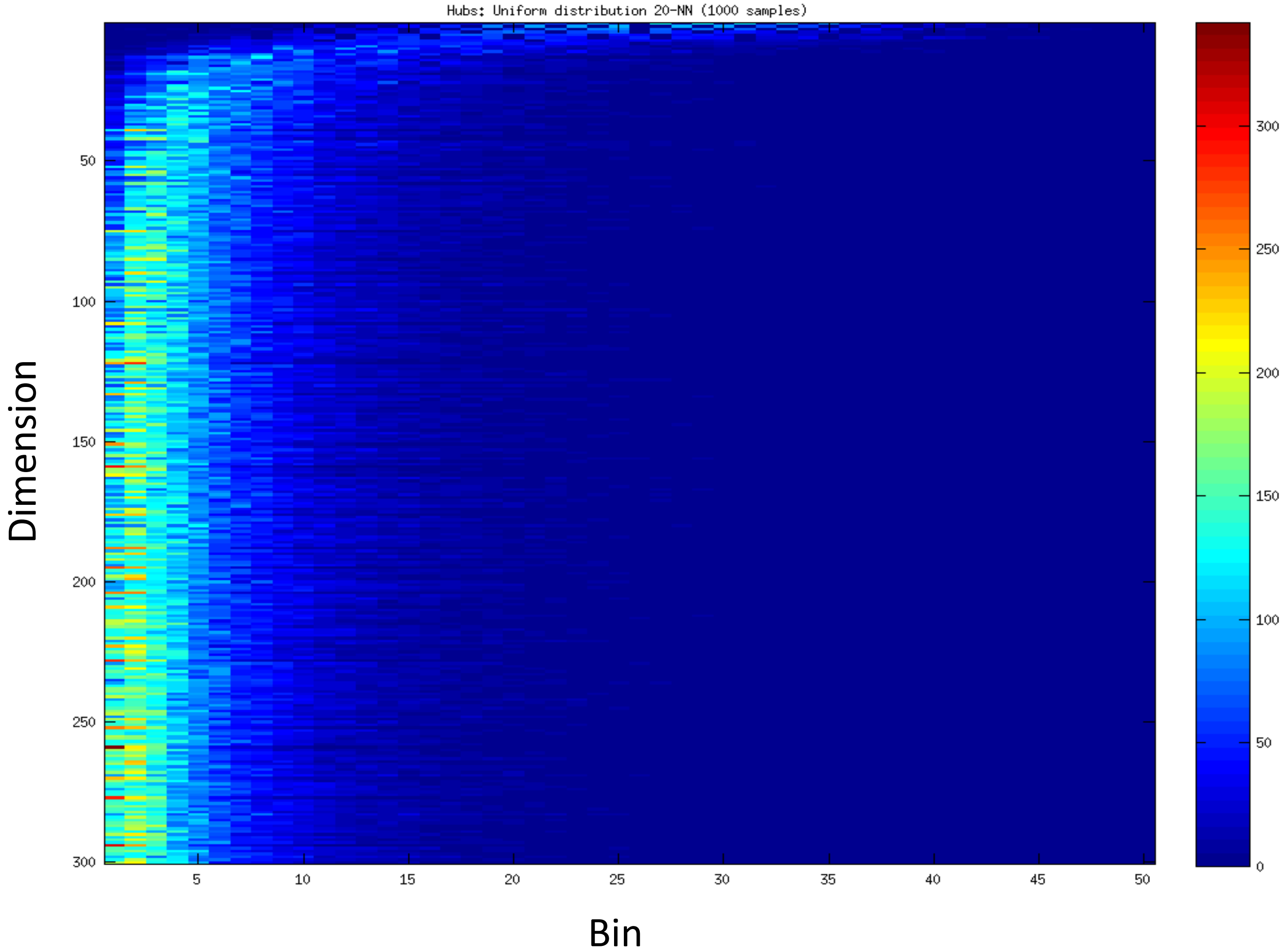
$$N_k(x) = \sum_i P_{ik}(x)$$

The distribution of N_k is skewed to the left. A small number of samples appear in the neighbourhood of many samples

20-NN M=100 (1000 samples) (50bins)



Hubness



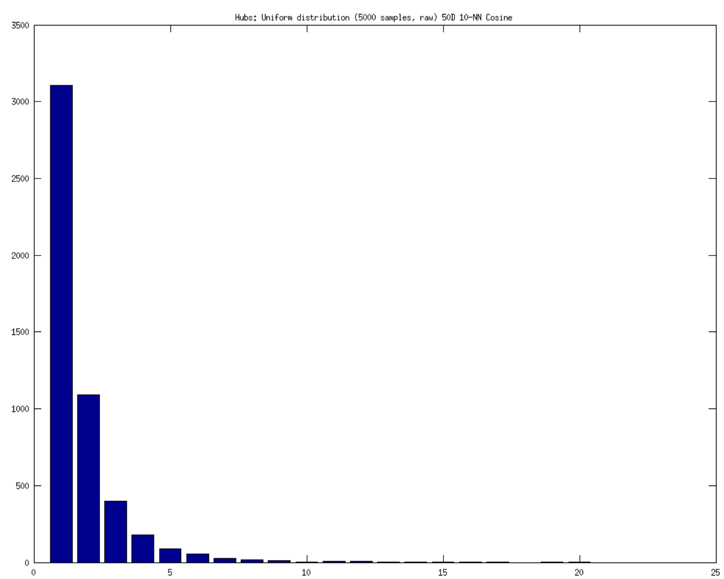
Hubs: Uniform distribution 20-NN (1000 samples)
Dimension
50
100
150
200
250
300
5
10
15
20
25
30
35
40
45
50
Bin
0
50
100
150
200
250
300

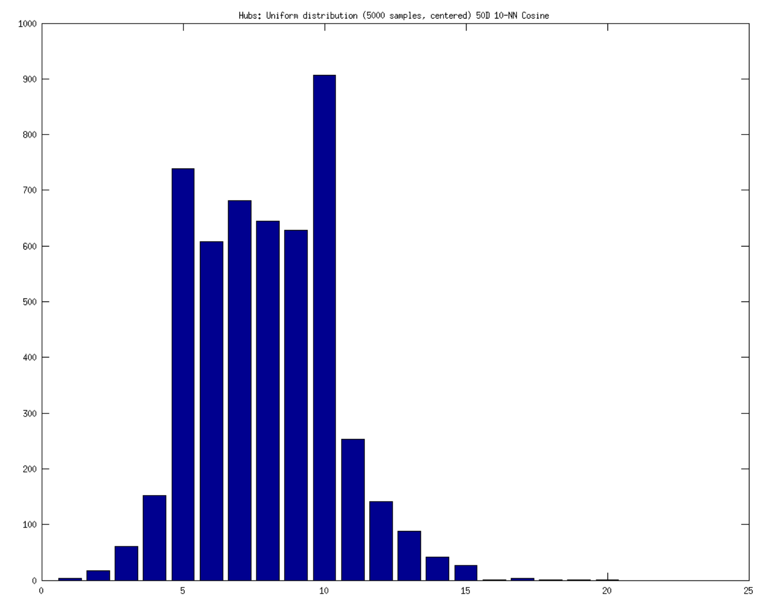
Hubs: centering

When using the cosine distance as similarity measure

$$d_{\cos}(x,y) = 1 - \frac{x^T y}{\|x\|\|y\|}$$

Centering the data helps reducing the hubness





I Suzuki et al. Centering Similarity Measures to Reduce Hubs.*2013 Conf. on Empirical Methods in NLP.*

Dimension reduction

- Space filling curves
- PCA
- FastMap
- IsoMap
- Random Projections (lemma)

SPACE FILLING CURVES

Space-filling curves

- Definition:
  - A continuous curve which passes through every point of a closed n-cell in Euclidean n-space E^n is called a *space filling curve (SFC)*.

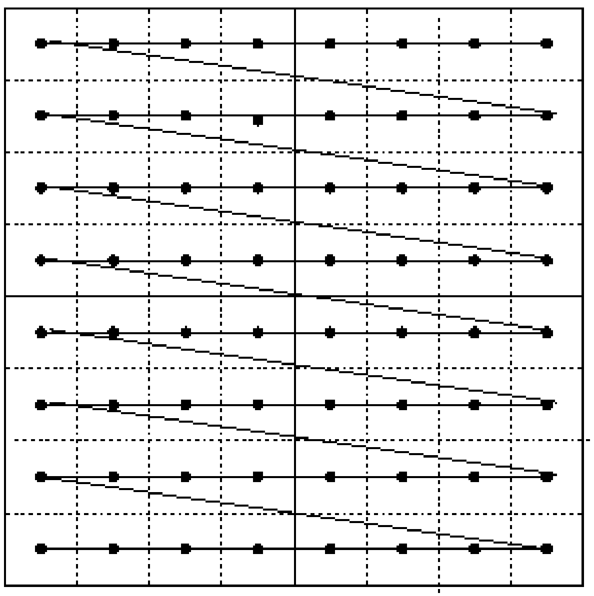
Application of SFC

- Mapping multi-dimensional space to one dimensional sequence
- Applications in computer science:
  - Database multi-attribute access
  - Image compression
  - Information visualization
  - ……

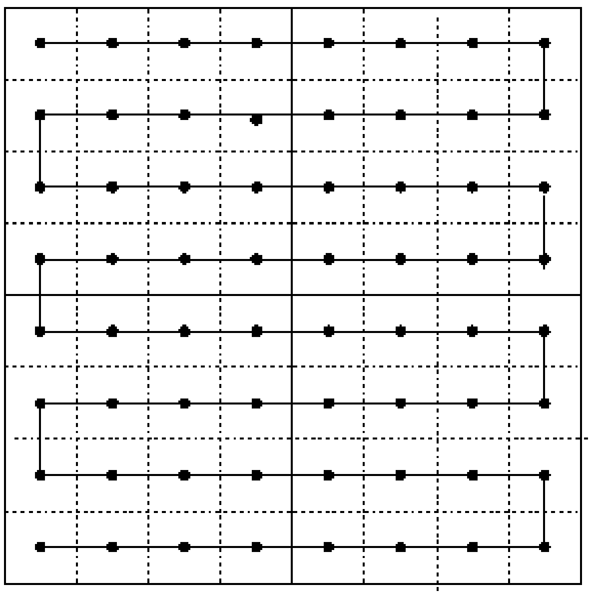
Categories of SFC

- Non-recursive
  - Z-Scan Curve
  - Snake Scan Curve
- Recursive
  - Hilbert Curve
  - Peano Curve
  - Gray Code Curve

Non-recursive Space Filling Curves



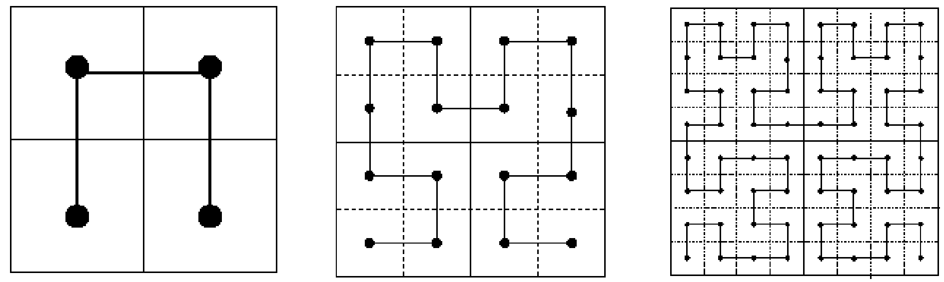
Z-Scan Curve



Snake Scan Curve

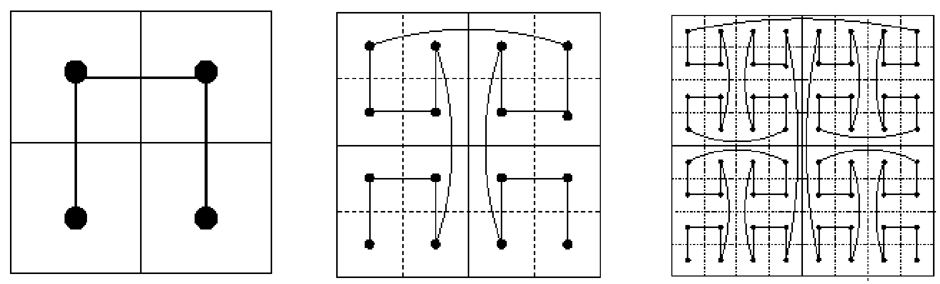
Recursive Space Filling Curves

Hilbert Curve



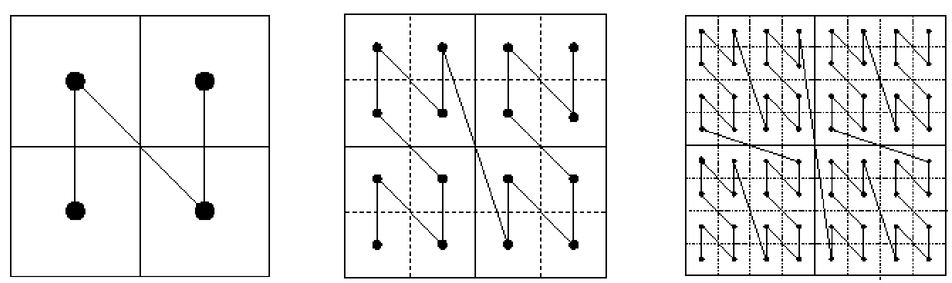
(a) (b) (c)

Gray Code Curve



(a) (b) (c)

Peano Curve



(a) (b) (c)

Properties of SFCs

- Coherent in Continuity
- Clustering Property

Z Curve → Hilbert Curve

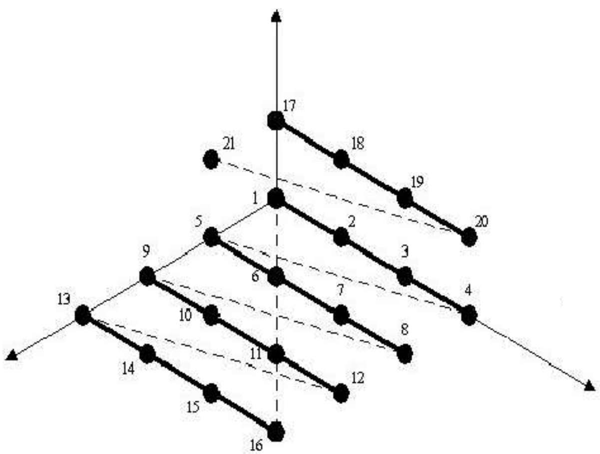
Good

- Direction Preserving

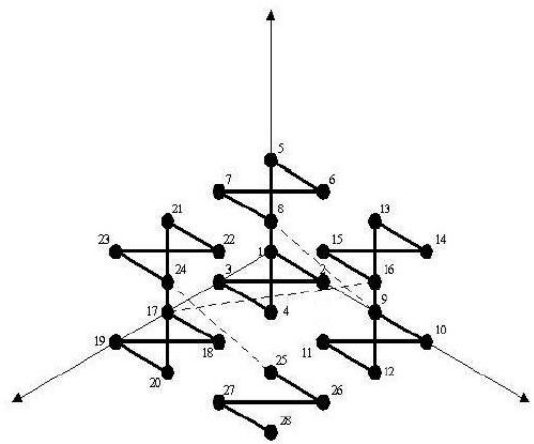
Hilbert Curve → Z Curve

Good

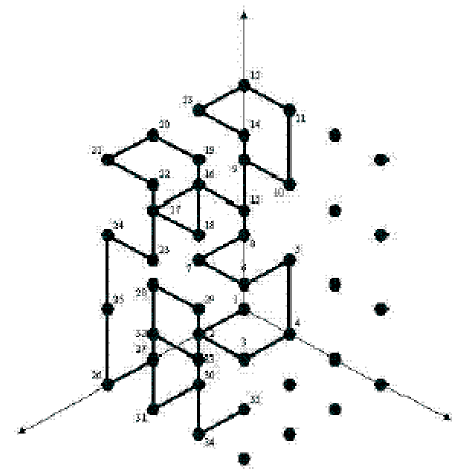
3D SFC



Z Curve



Peano Curve



Hilbert Curve

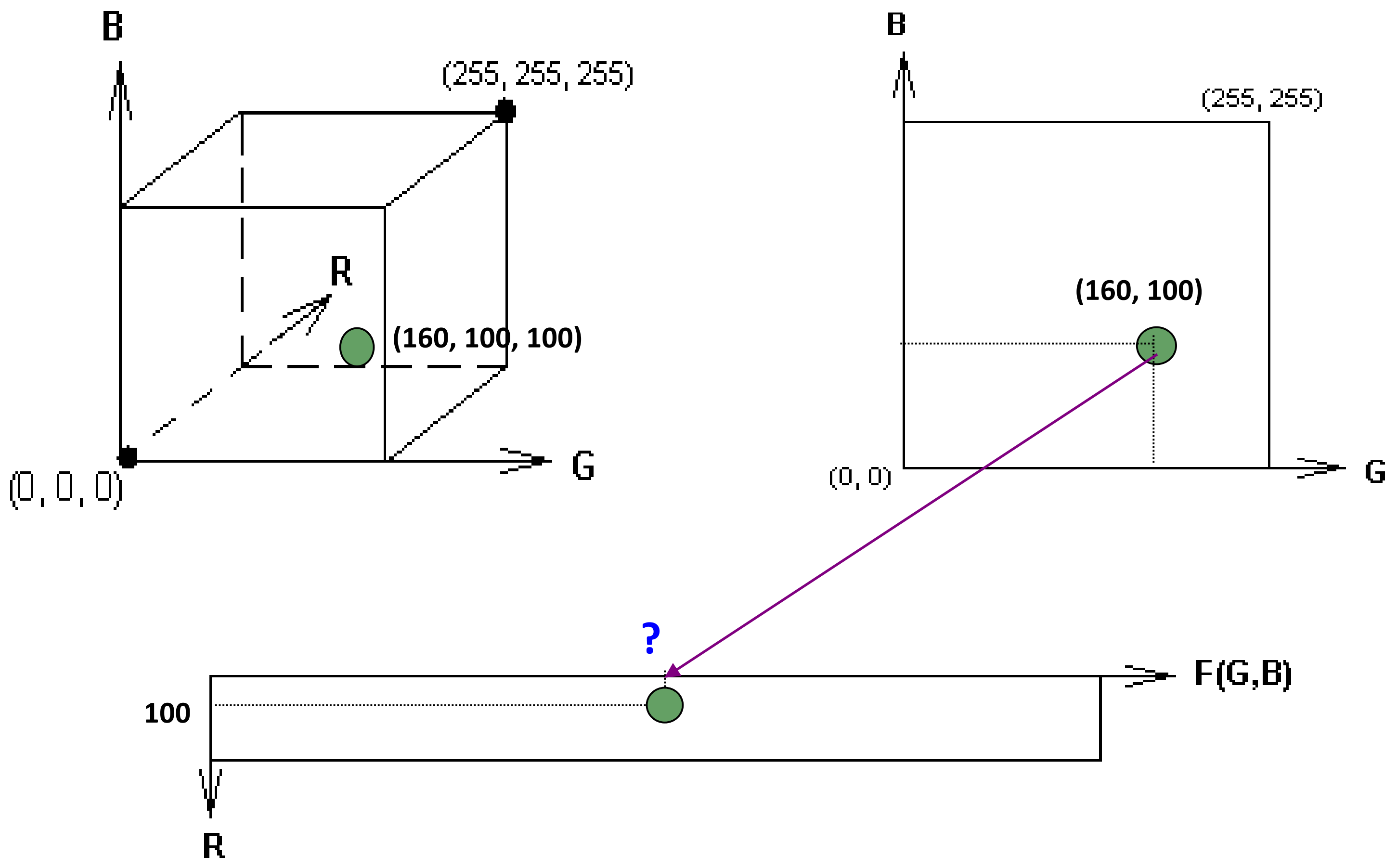
N-dimensional algorithm:

A.R. Butz (April 1971). "Alternative algorithm for Hilbert's space filling curve.". *IEEE Trans. On Computers,* **20**: 424–42.

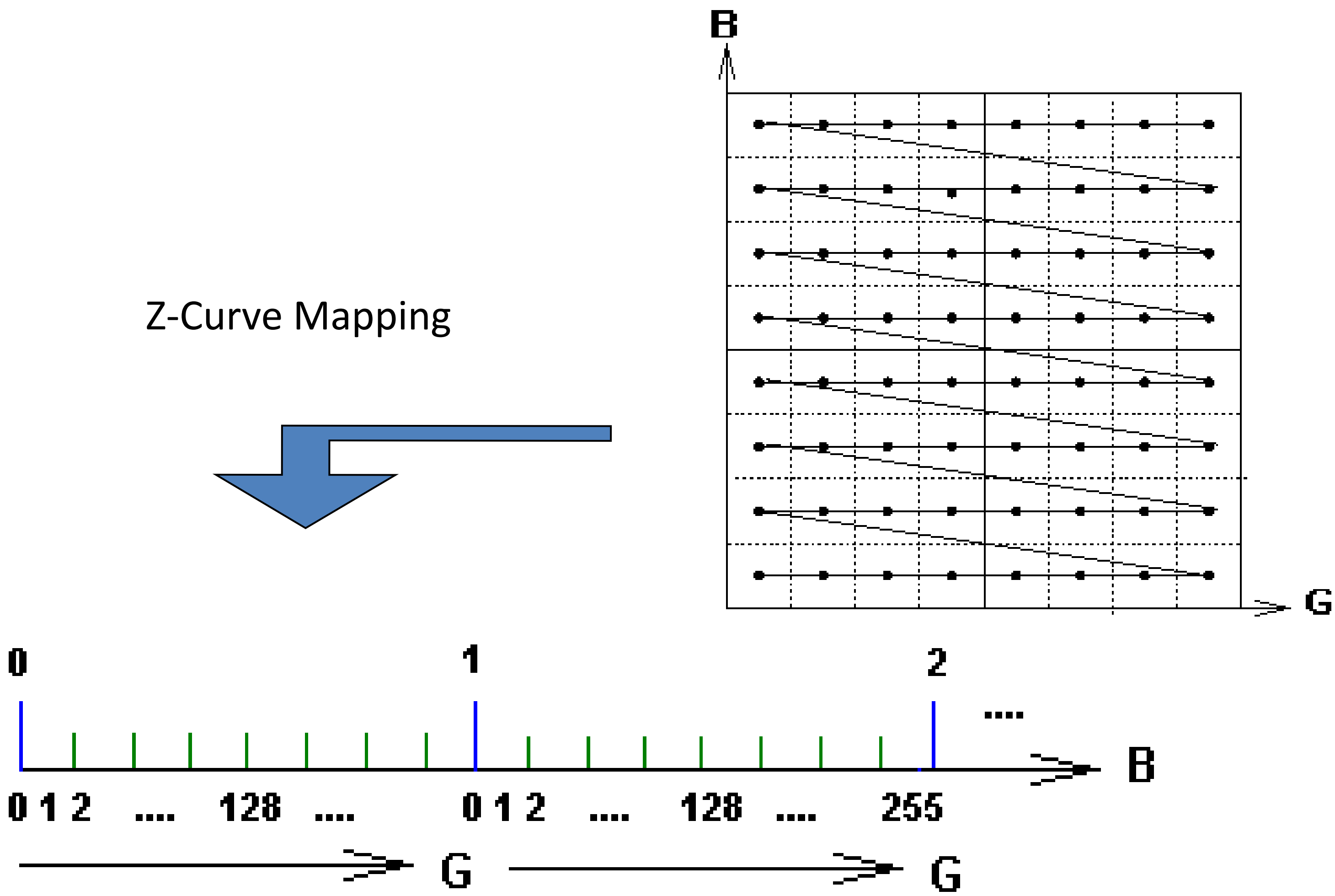
SFC in Information Visualization

- Example Data
  - Color (R, G, B: [0, 255])
  - Data with obvious geometric pattern
    - 4D Hyper Sphere
  - Data without obvious geometric pattern
    - Iris flowers (5 attributes, 3 classes)
- Example SFC
  - Z-Curve
  - Hilbert Curve

Visualizing Color (RGB)

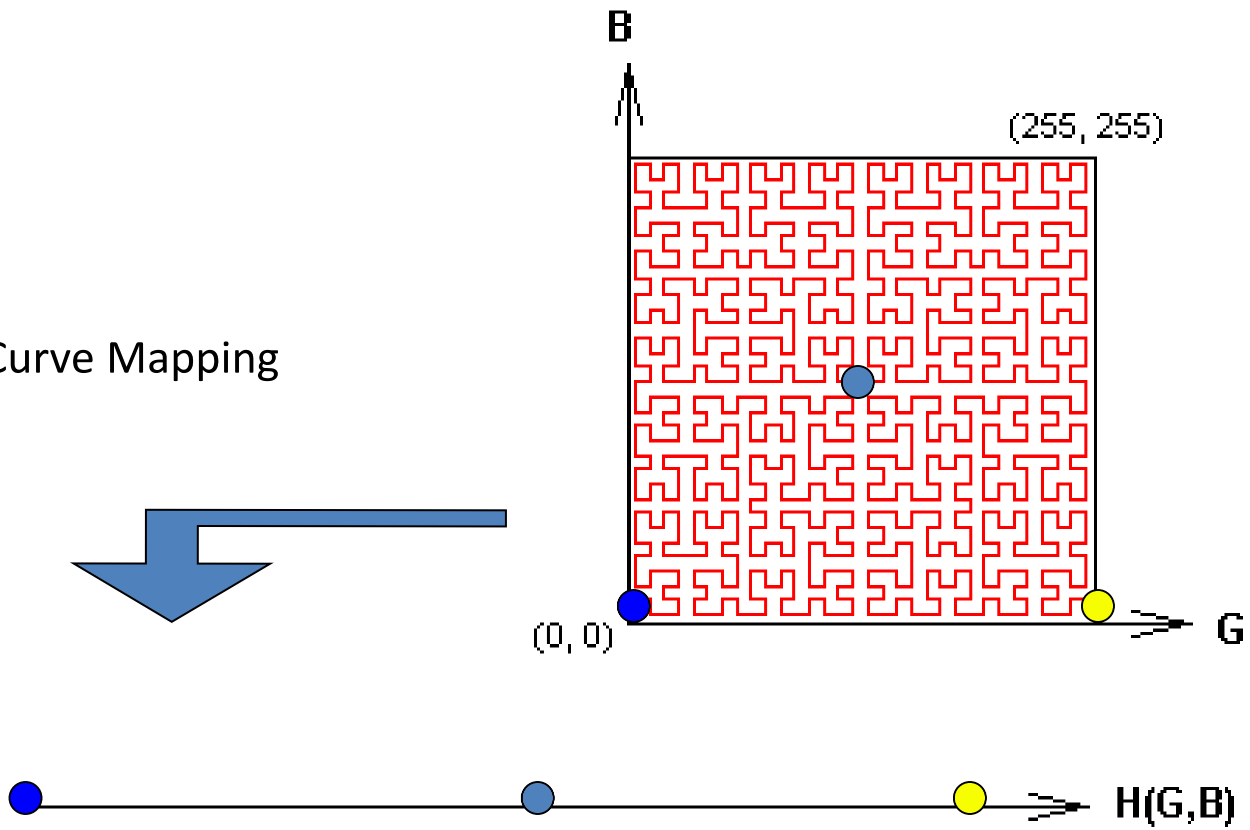


Visualizing Color (RGB)

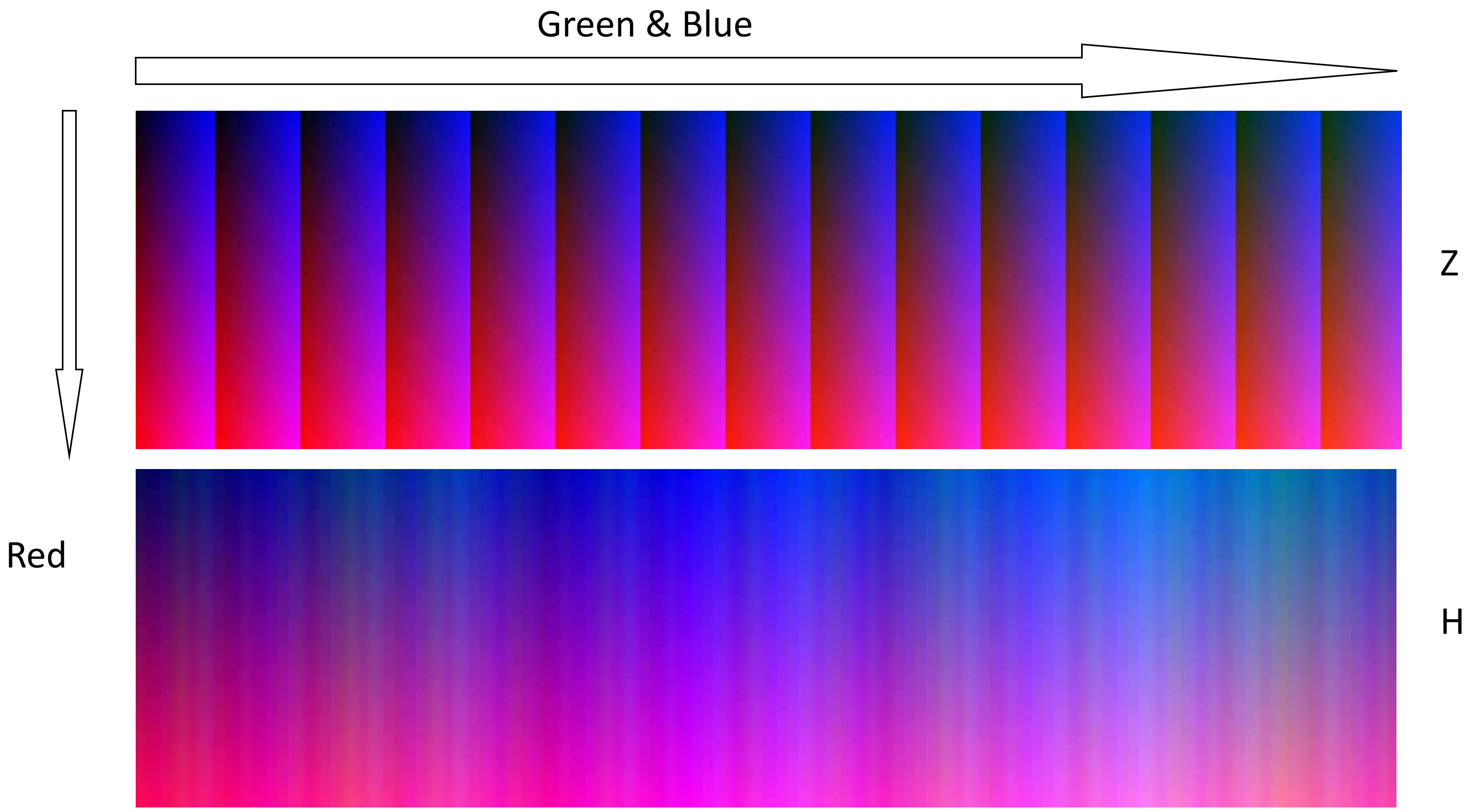


Visualizing Color (RGB)

Hilbert Curve Mapping

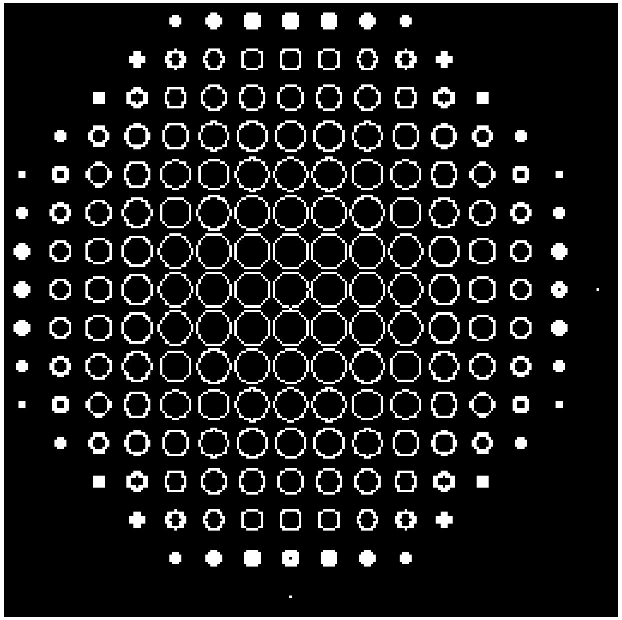


Visualizing Color (RGB)

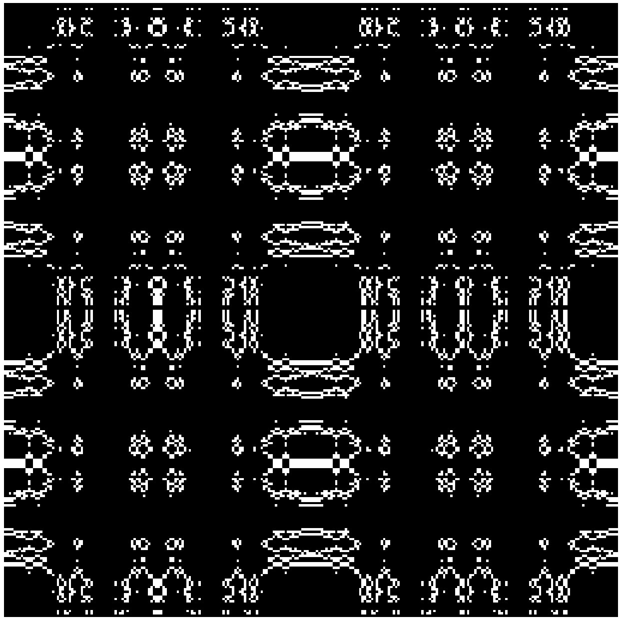


Visualizing 4D Hyper-Sphere Surface

- Z-Curve

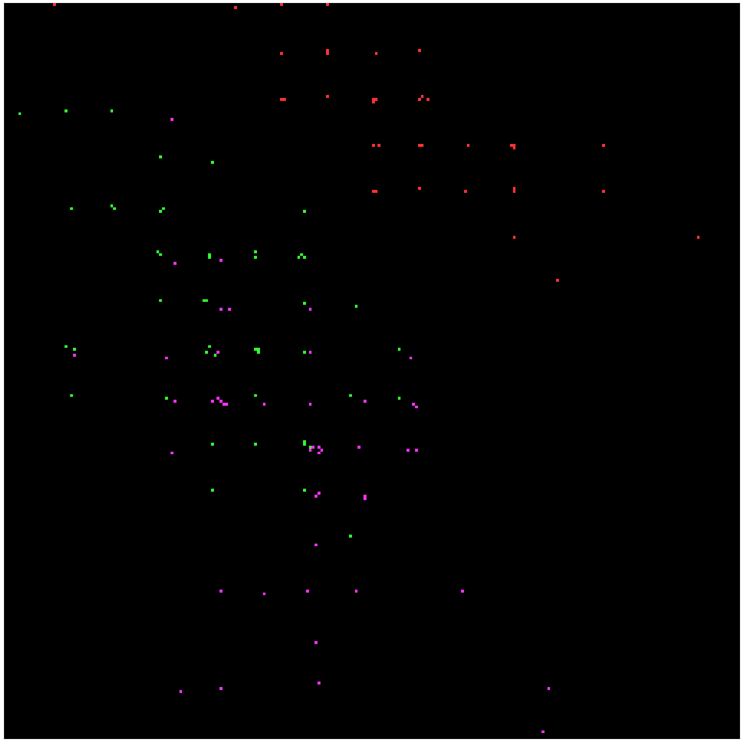


- Hilbert Curve

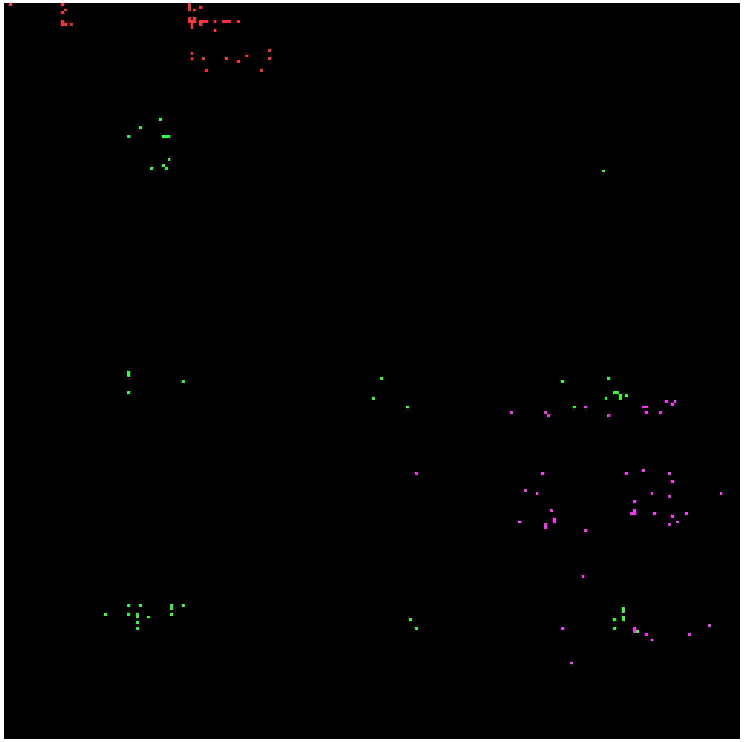


The Visualization of Iris Flowers

- Z-Curve



- Hilbert Curve



The Visualization of Iris Flowers

- Extended Hilbert Curve