Random Projections

MDS : d_{ij} $i=1,\dots,n$

$Y_i \in \mathbb{R}^k$

$$\min_{Y \in \mathbb{R}^{k \times n}} \sum_{i,j} \left(\|Y_i - Y_j\|^2 - d_{ij}^2 \right)^2$$

Total square error distortion

"Average"

$$d_{ij}(1-\varepsilon) \le \|Y_i - Y_j\| \le d_{ij}(1+\varepsilon) \quad \forall\, i,j$$

Uniform distortion

① SDP $\|Y_i - Y_j\|^2 \to$ SD. Relaxation

② Universal basis 不依赖数据

Random Projection

$X_i \in \mathbb{R}^p$. $d_{ij} = \|X_i - X_j\|$ $i=1,\dots,n$.

$Y_i = f(X_i) \in \mathbb{R}^k$ $k = O(c(\varepsilon)\log n)$

s.t.

$$1-\varepsilon \le \frac{\|Y_i - Y_j\|}{\|X_i - X_j\|} \le 1+\varepsilon$$

with probability $\ge 1 - n^{-\alpha}$, $\alpha > 0$ $\to 1$

f : random proj.

Almost Isometry Embedding!

Johnson–Lindenstrauss Lemma

John. Lindenstrauss

Lipschitz Extension

概率性方法证明存在性

2001 2003 Computer Science 数据压缩

Approximate NN.

Sanjoy Dasgupta, Anupam Gupta

Dimitris Achlioptas '03

Thm Given $1>\varepsilon>0$, n, $\alpha>0$, Let

$$k = C(\alpha,\varepsilon)\log n = (4+2\alpha)\left(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}\right)^{-1}\log n.$$

Then for any n points $x_i \in \mathbb{R}^d$ $(i=1,\dots,n)$, there exists a map $f:\mathbb{R}^d\to\mathbb{R}^k$ such that $\forall x_i, x_j$

$$1-\epsilon \le \frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \le (1+\epsilon) \qquad (*)$$

f can be founded in randomized polynomial time. $(*)$ holds with probability at least $1-n^{-\alpha}$, $\alpha>0$.

例: 1) f can be constructed by random matrices/projections

$f(x_i) = R x_i$ $\qquad X = [x_1 \cdots x_n]^{d\times n}$

$\Rightarrow Y^{k\times n} = R^{k\times d} X^{d\times n}$ $\qquad Y=[y_1 \dots y_n]^{k\times n}$

- $R=[r_1,\cdots,r_k]^T$ $\quad r_i\in S^{d-1}$, e.g. $r_i = \frac{(a_1^i \cdots a_d^i)}{\|a^i\|}$ $\quad a_d^i \sim N(0,1)$
- $R = A/\sqrt{k}$ $\quad A_{ij}\sim N(0,1)$, variance $=1$
- $R = A/\sqrt{k}$ $\quad A_{ij} = \begin{cases} 1 & p=\frac{1}{2} \\ -1 & (1-p)=\frac{1}{2}\end{cases}$, $\quad A_{ij} = \begin{cases} 1 & p=1/6 \\ 0 & 2/3 \\ -1 & p=1/6\end{cases}$

Human Immigration

640k   5k 640k.   100k

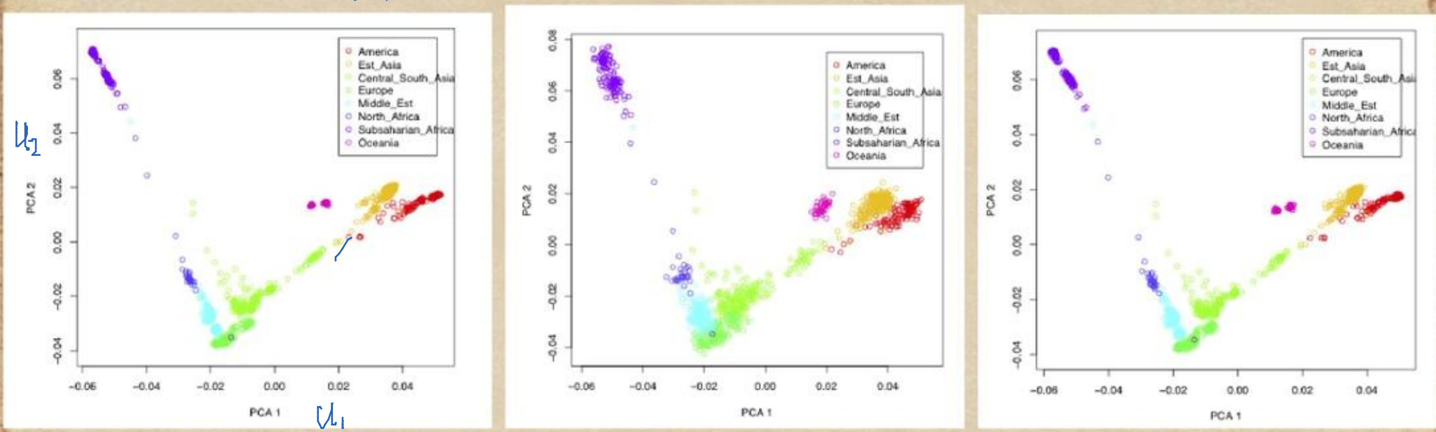

FIGURE 1. (Left) Projection of 1043 individuals on the top 2 MDS principal components. (Middle) MDS computed from 5,000 random projections. (Right) MDS computed from 100,000 random projections. Pictures are due to Qing Wang.

14年10月7日星期二

SNPs. 分析 (Single Nucleid Polymorphisms)

600k.

Human Genome Diversity Project $n=1064$ individual

$p=644,258$ SNPs $X^{n\times p}$

0: AA   1: AC   2: CC   9: Missing value.   ~21 人.

$X^{n\times p}$ $n=1043$, $p=644,258$

SVD. $\tilde{X} = HX$ $H = (I - \frac{1}{n} 1 1^T)$

$\tilde{X} = U \overset{p\times n}{\Sigma} V^{T} \cdot \underline{p\times p}$

取U_1 -- 取U_k PCA. $n=1043$ 人

$\mathbb{R}^d$ uniformly projected to k-subspace

$\overset{\text{dist}}{\Longleftrightarrow}$ random vector on S^{d-1} restricted onto top k-coord.

$$(\underbrace{a_1^i \cdots}_{k} a_d^i) \qquad \|a^i\|_2 = 1$$

$$(\underbrace{a_1^i \cdots a_k^i}, 0)$$

$x_i \sim N(0,1)$, $i = 1, \ldots, d$.

$X = (x_1 \cdots x_d)$ $\quad Y = \frac{X}{\|X\|} \in S^{d-1}$ uniformly distributed.

$Z = (x_1, \cdots, x_k, 0 \cdots 0) \in \mathbb{R}^d$ $\quad$ k-subspace

$L := \|Z\|^2$ $\quad \mathbb{E}L = \frac{k}{d} = \mu$

Lemma Concentration Inequality $\quad k < d$.

(a) $\beta < 1$ lower bound.

$$\text{Prob}[L \le \beta\mu] \le \beta^{k/2}\left(1 + \frac{(1-\beta)k}{d-k}\right)^{d-k/2} \le \exp\left(\frac{k}{2}(1-\beta+\ln\beta)\right)$$

(b) $\beta > 1$ upper bound

$$\text{Prob}[L \ge \beta\mu] \le \beta^{k/2}\left(1 + \frac{(1-\beta)k}{d-k}\right)^{d-k/2} \le \exp\left(\frac{k}{2}(1-\beta+\ln\beta)\right)$$

$\mathbb{E}L = \mu$. exponentially $\sim \mu$ $\checkmark$

Proof of J-L Lemma

$d < k$, trivial

$d > k$.

k-subspace $\quad x_i \in \mathbb{R}^d \to y_i \in \mathbb{R}^k \quad i = 1, \dots, n.$

$L = \|y_i - y_j\|^2 \qquad \mu = \mathbb{E}L = \frac{k}{d}\|x_i - x_j\|^2 \qquad y_i = (x_1^i \cdots x_k^i\ 0)$

$$\text{Prob}[L \le (1-\varepsilon)\mu] \le \exp\left(\frac{k}{2}\left(1 - (1-\varepsilon) + \ln(1-\varepsilon)\right)\right)$$

β

$\ln(1-\varepsilon) \le -\varepsilon - \varepsilon^2/2 \qquad \varepsilon \in [0,1)$

$$\le \exp\left[\frac{k}{2}\left(\varepsilon - \left(\varepsilon + \frac{\varepsilon^2}{2}\right)\right)\right] = \exp\left(-\frac{k\varepsilon^2}{4}\right)$$

$$\le \exp\left(-(2+\alpha)\ln n\right) \qquad k \ge 4(1+\alpha/2)(\varepsilon^2/2)^{-1}\ln n$$

$$= n^{-(2+\alpha)} \to 0$$

$$\text{Prob}[L \ge (1+\varepsilon)\mu] \le \exp\left[\frac{k}{2}\left(1 - (1+\varepsilon) + \ln(1+\varepsilon)\right)\right]$$

$$\le \exp\left[\frac{k}{2}\left(\varepsilon^2/2 - \varepsilon^3/3\right)\right],$$

$\ln(1+\varepsilon) \le \varepsilon - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{3}, \ \varepsilon \ge 0$

$$\le \exp\left(-(2+\alpha)\ln n\right) = n^{-(2+\alpha)} \to 0$$

$k \ge 4(1+\alpha/2)(\varepsilon^2/2 - \varepsilon^3/3)^{-1}\ln n$

Given pair (i, j).

$$\text{Prob}\left\{1-\varepsilon \le \frac{\|y_i - y_j\|^2}{\|x_i - x_j\|^2} \le 1+\varepsilon\right\} \le 1 - \frac{2}{n^{2+\alpha}}$$

$\forall (i,j)$ from $\{i=1,\dots,n;\ j=1,\dots,n\}$ of $\binom{n}{2}$

$$\text{Prob}\left\{\forall (i,j),\ 1-\varepsilon \le \frac{\|y_i - y_j\|^2}{\|x_i - x_j\|^2} \le 1+\varepsilon\right\} \le 1 - \binom{n}{2}\frac{2}{n^{2+\alpha}} \le 1 - n^{-\alpha}$$

$\binom{n}{2} = n(n-1)/2$