UL = RANDOM PROJECTIONS

Random projections provide an alternative to PCA for dimension reduction.

The problem: given a learning sample $\mathcal{L}_n = \{x_1, \dots, x_n\}$ of size n, $x_i \in \mathbb{R}^D$, D supposed very large, project these points into a lower dimensional space, of dimension $d \ll D$, preserving some information.

↳ we are interested in preserving the geometry of $x_1, \dots, x_n$.

More precisely, our goal is to construct a mapping

$T : \mathbb{R}^D \to \mathbb{R}^d$ (possibly random, and it will be)

such that

$$\mathbb{P}\left(\forall i \neq j : 1-\varepsilon \leq \frac{\|T(x_i) - T(x_j)\|_2}{\|x_i - x_j\|_2} \leq 1+\varepsilon \right) \geq 1-\delta,$$

$\varepsilon > 0$ small
$\delta \in (0,1)$

↑ With high probability, points that are close to each other stay close

different philosophy than PCA, which is projecting the data in directions of maximum variance.

We will see shortly that this is possible with a simple random linear transformation T. We recall first some background results about sub-Gaussian & sub-Gamma distributions.

I- PRELIMINARIES

I.1. Sub-Gaussian distributions.

- Let $X \sim \mathcal{N}(\mu, \sigma^2)$. The Moment Generating Function (MGF) of X is given by $M_X(\theta) := \mathbb{E}\{e^{\theta X}\}$, with

(*) $$\log\left[\mathbb{E}\left\{e^{\theta(X-\mu)}\right\}\right] = \frac{\theta^2\sigma^2}{2}, \quad \forall \theta \in \mathbb{R}.$$

↳ Useful to obtain bounds on tail probabilities of a distribution:

CHERNOFF'S BOUND: For a real-value RV X,
$\forall t \in \mathbb{R}$, put $\Lambda_X(t) := \sup_{\theta \geq 0} (\theta t - \log M_X(\theta))$
Then $\mathbb{P}(X > t) \leq \exp(-\Lambda_X(t))$

Indeed, $\mathbb{P}(X > t) = \mathbb{P}(e^{\theta X} > e^{\theta t})$
$\leq e^{-\theta t}\, \mathbb{E}(e^{\theta X})$ ← Markov
$= e^{-\theta t}\, M_X(\theta)$

For $X \sim \mathcal{N}(\mu, \sigma^2)$, we immediately get

$$\left.\begin{array}{l} \mathbb{P}(X-\mu > t) \\ \mathbb{P}(X-\mu < -t) \end{array}\right\} \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

- We introduce a class of distributions that have lighter tails than the Gaussian distribution:

A RV X is said to be SUB-GAUSSIAN if $\exists s^2 > 0$ s.t. $\forall \theta \in \mathbb{R}$, $\log\left[\mathbb{E}\{e^{\theta(X-\mathbb{E}X)}\}\right] \leq \frac{\theta^2 s^2}{2}$.

Whenever this holds, we write $X \in SG(s^2)$.
The smallest $s^2 > 0$ for which $X \in SG(s^2)$ is called the VARIANCE PROXY of X, and is denoted $\|X\|_{vp}^2$.

It is given by

$$\|X\|_{vp}^2 = \sup_{\theta \neq 0} \frac{2}{\theta^2} \log \mathbb{E}\left\{ e^{\theta(X - \mathbb{E}X)} \right\}.$$

As the notation indicates, the variance proxy behaves almost like a norm: for two sub-Gaussian RVs X, Y, $\forall \alpha \in \mathbb{R}$,

(i) $\|\alpha X\|_{vp} = |\alpha| \, \|X\|_{vp}$

(ii) $\|X + Y\|_{vp} \leq \|X\|_{vp} + \|Y\|_{vp}$

(iii) $\|X\|_{vp} = 0 \iff X = c$ a.s.

(i) follows from the definition of $\|X\|_{vp}^2$

(ii) $\forall 1 < p, q < +\infty$ s.t. $p^{-1} + q^{-1} = 1$, X, Y centered,

$$\mathbb{E}\left\{ e^{\theta(X+Y)} \right\} = \mathbb{E}\left\{ e^{\theta X} e^{\theta Y} \right\} \quad \text{(Hölder)}$$

$$\leq \left(\mathbb{E}\left\{ e^{p\theta X} \right\} \right)^{1/p} \left(\mathbb{E}\left\{ e^{q\theta Y} \right\} \right)^{1/q}$$

$$\leq \left(\exp\left\{ \frac{p^2 \theta^2 \|X\|_{vp}^2}{2} \right\} \right)^{1/p} \left(\exp\left\{ \frac{q^2 \theta^2 \|Y\|_{vp}^2}{2} \right\} \right)^{1/q}$$

$$= \exp\left\{ \frac{\theta^2}{2} \left(p \|X\|_{vp}^2 + q \|Y\|_{vp}^2 \right) \right\}$$

$$= \exp\left\{ \frac{\theta^2}{2} \left(p \|X\|_{vp}^2 + \frac{p}{p-1} \|Y\|_{vp}^2 \right) \right\}$$

Holds for any $p > 1$.

$$\inf_{p>1} \left\{ p \|X\|_{vp}^2 + \frac{p}{p-1} \|Y\|_{vp}^2 \right\} = \left(\|X\|_{vp} + \|Y\|_{vp} \right)^2 \quad \text{achieved for } p = 1 + \frac{\|Y\|_{vp}}{\|X\|_{vp}}$$

Thus $\mathbb{E}\left\{ e^{\theta(X+Y)} \right\} \leq \exp\left\{ \frac{\theta^2}{2} \left(\|X\|_{vp} + \|Y\|_{vp} \right)^2 \right\}$

(iii) Note that $\forall \theta \in \mathbb{R}$ $\quad 1 = e^{\mathbb{E}(\theta X)} \leq \mathbb{E}\left(e^{\theta X} \right)$

(X assumed centered)

In addition, $\|X\|_{vp} = 0 \Rightarrow \mathbb{E}\left(e^{\theta X} \right) \leq 1$.

(X sub-gauss)

$\Rightarrow e^{\mathbb{E}(\theta X)} = \mathbb{E}\left(e^{\theta X} \right)$, which can happen only if X = constant a.s. (here equal to 0)

Theorem: Suppose X sub-gaussian. Then $\forall t \in \mathbb{R}$,

$$\left.\begin{matrix} \mathbb{P}(X - \mathbb{E}X > t) \\ \mathbb{P}(X - \mathbb{E}X < -t) \end{matrix}\right\} \leq \exp\left(- \frac{t^2}{2\|X\|_{vp}^2} \right).$$

proof = obvious ■

(X is sub gaussian $\iff -X$ is sub gaussian.)

× Ex: Any bounded RV is sub-Gaussian:

Let $X \in [a, b]$ a.s.

Then $\forall \theta \in \mathbb{R}$, $\log\left[\mathbb{E}\left\{ e^{\theta(X - \mathbb{E}X)} \right\} \right] \leq \frac{\theta^2 (b-a)^2}{8}$

$\Rightarrow \|X\|_{vp}^2 \leq \frac{(b-a)^2}{4}$. (not obvious – proof required)

Remark = Any Gaussian RV is sub-Gaussian with variance proxy equal to its variance. However, in general, any sub-Gaussian RV X has variance smaller or equal to its variance proxy: $\mathrm{Var}\, X \leq \|X\|_{vp}^2$.

I.2. Sub-gamma distributions.

Recall that for $X \sim \gamma(a,b)$, $a,b>0$,

$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-\beta x}, \quad x \geq 0.$$

$$\mathbb{E}\left\{e^{\theta(X-\mathbb{E}X)}\right\} = \begin{cases} \left(\frac{b}{b-\theta}\right)^a e^{-\frac{a\theta}{b}} & \text{if } \theta < b \\ +\infty & \text{if } \theta \geq b \end{cases}$$

Using $-\log(1-u)-u \leq \frac{u^2}{2(1-u)}$, $\forall u \in (0,1)$,

we have $\forall \theta < b$,

$$\log\left[\mathbb{E}\left\{e^{\theta(X-\mathbb{E}X)}\right\}\right] = -a\log\left(1-\frac{\theta}{b}\right) - \frac{a\theta}{b} \leq \frac{\theta^2\alpha^2}{2(1-\theta\beta)} \quad \text{with} \quad \alpha = \frac{\sqrt{a}}{b}, \quad \beta = \frac{1}{b}$$

leads to ↓

Def = A real-valued RV X is said to be sub-gamma on the right tail if $\exists \alpha>0$, $\beta>0$ s.t.

$$\forall \theta < \frac{1}{\beta} \quad \log\left[\mathbb{E}\left\{e^{\theta(X-\mathbb{E}X)}\right\}\right] \leq \frac{\theta^2\alpha^2}{2(1-\theta\beta)}$$

and we write $X \in \Gamma_+(\alpha,\beta)$

X is said to be sub-gamma in the left tail if

$$\forall \theta > -\frac{1}{\beta} \quad \log\left[\mathbb{E}\left\{e^{\theta(X-\mathbb{E}X)}\right\}\right] \leq \frac{\theta^2\alpha^2}{2(1-\theta\beta)},$$

and we write $X \in \Gamma_-(\alpha,\beta)$

$X \in \Gamma_+(\alpha,\beta) \Leftrightarrow -X \in \Gamma_-(\alpha,\beta)$.

We say that X is sub-gamma, and we write $X \in \Gamma(\alpha,\beta)$ if $X \in \Gamma_+(\alpha,\beta)$ and $X \in \Gamma_-(\alpha,\beta)$.

→ From previous calculations, we see that for $a,b>0$, $X \sim \gamma(a,b)$

$$\Rightarrow X \in \Gamma_+\left(\frac{\sqrt{a}}{b}, \frac{1}{b}\right)$$

$$-X \in \Gamma_-\left(\frac{\sqrt{a}}{b}, \frac{1}{b}\right).$$

× Remark: By definition, any sub-gaussian variable X is sub-gamma: $X \in \Gamma(\|X\|_{\psi_2}, \beta)$ $\forall \beta>0$. Moreover, if X is sub-gaussian, then $X^2 \in \Gamma(4e\|X\|_{\psi_2}^2, 4e\|X\|_{\psi_2}^2)$ (proof required)

We finish this section with a concentration inequality for sub-gamma RVs:

Theorem: Let $X_1, \dots, X_n$ be independent sub-gamma RVs, $X_i \in \Gamma(a_i, b_i)$. Then, $\forall \varepsilon \in (0,1)$, $\forall t>0$,

$$\left.\begin{array}{l} \mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) > t\right) \\ \mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) < -t\right) \end{array}\right\} \leq e^{-\frac{t}{2}\left(\frac{\varepsilon}{\max b_i} \wedge \frac{(1-\varepsilon)t}{\sum a_i^2}\right)}$$

Proof of this theorem + proofs of other results presented here can be found in the teaching material of Quentin Paris:
www.qparis-math.com/teaching [High dimensional statistical methods]

II - JOHNSON - LINDENSTRAUSS THEOREM.

We are now in a position to solve the problem stated in the introduction, and to prove the following result, known as the Johnson-Lindenstrauss Theorem.

JOHNSON-LINDENSTRAUSS THEOREM

Let $x_1, \dots, x_n$ = n distinct (non-random) points in $\mathbb{R}^D$.
Fix $\varepsilon > 0$, $\delta \in (0,1)$.
Let M be a $d \times D$ random matrix, whose rows $r_1, \dots, r_d$ are
- → independent
- → centered
- → isotropic $\mathbb{E}(r_i r_i^t) = I_D$
- → sub-gaussian with variance proxy σ^2.
  ↳ $u^t r_i$ is sub-gaussian $\forall u \in \mathbb{R}^D$

Put $T := \frac{1}{\sqrt{d}} M$.

Then, provided $d \gtrsim \frac{\sigma^4}{\varepsilon^2} \log\left(\frac{n}{\sqrt{\delta}}\right)$,

↑ constant times ...
(constant is explicit - see proof of theorem)

we have

$$\mathbb{P}\left(\forall i \neq j : 1-\varepsilon \leq \frac{\|T(x_i) - T(x_j)\|_2}{\|x_i - x_j\|_2} \leq 1+\varepsilon\right) \geq 1-\delta$$

↑ T is very easy to implement.
Take for example a matrix $M \in \mathbb{R}^{d \times D}$ with iid entries $\mathcal{N}(0, \sigma^2)$, or any distribution with bounded support.

proof = Denote $\mathcal{X} := \{x_1, \dots, x_n\}$.

By linearity of T,

$$\forall x, y \in \mathcal{X}, \quad 1-\varepsilon \leq \frac{\|T(x) - T(y)\|_2}{\|x-y\|_2} \leq 1+\varepsilon$$

$$\Leftrightarrow \quad 1-\varepsilon \leq \|T(z)\|_2 \leq 1+\varepsilon$$

with $z := \frac{x-y}{\|x-y\|_2}$

Since $(1-\varepsilon)^2 \leq 1-\varepsilon$, and $1+\varepsilon \leq (1+\varepsilon)^2$, the last written expression is equivalent to $1-\varepsilon \leq \|T(z)\|_2^2 \leq 1+\varepsilon$, and so it is enough to show that

$$\mathbb{P}\left(\forall z \in \mathcal{Z} : 1-\varepsilon \leq \|T(z)\|_2^2 \leq 1+\varepsilon\right) \geq 1-\delta, \quad (*)$$

where $\mathcal{Z} := \left\{ \frac{x-y}{\|x-y\|_2}, \; x \neq y \in \mathcal{X} \right\}$.

By construction, $T = \underset{(d \times D)}{\frac{1}{\sqrt{d}} M} = \frac{1}{\sqrt{d}} \begin{pmatrix} — \; r_1 \; — \\ — \; r_d \; — \end{pmatrix}$ (d rows, D columns),

so that $\|T(z)\|_2^2 = \frac{1}{d} \sum_{i=1}^{d} r_i^t z$

↑ Also denoted $\langle r_i, z \rangle$.

Note that z is a unit vector.

+ r_i is sub-gaussian with variance proxy σ^2
  (= $\max_{\|u\|_2=1} \|u^t r_i\|_{\psi_2}^2$ by def of the variance proxy of a random vector)

$\Rightarrow \langle r_i, z \rangle \in SG(\sigma^2)$

& $\langle r_i, z \rangle^2 \in \Gamma(4e\sigma^2, 4e\sigma^2)$.

In addition,

$$1-\varepsilon \leqslant \|T(z)\|_2^2 \leqslant 1+\varepsilon \iff \left| \frac{1}{d} \sum_{i=1}^{d} \langle r_i, z\rangle^2 - 1 \right| \leqslant \varepsilon$$

&

$$\begin{aligned}\mathbb{E}\langle r_i, z\rangle^2 &= \mathbb{E}\left((r_i^t z)(r_i^t z) \right) \\ &= \mathbb{E}\left(z^t r_i r_i^t z \right) \\ &= z^t \underbrace{\left(\mathbb{E}\, r_i r_i^t\right)}_{\equiv I_D} z = z^t z = 1.\end{aligned}$$

Thus,

$$\left| \frac{1}{d} \sum_{i=1}^{d} \langle r_i, z\rangle^2 - 1 \right| = \frac{1}{d} \left| \sum_{i=1}^{d} \left(\langle r_i, z\rangle^2 - \mathbb{E}\langle r_i, z\rangle^2 \right) \right|,$$

and so it is enough to show that (see (*) page 8)

$$(**) \quad \mathbb{P}\left(\max_{z \in \mathcal{Z}} \left| \sum_{i=1}^{d} Y_i(z) - \mathbb{E}\, Y_i(z) \right| \leqslant d\varepsilon \right) \geqslant 1-\delta,$$

where

$$Y_i(z) := \langle r_i, z\rangle^2 \in \Gamma(4e\sigma^2, 4e\sigma^2).$$

$$\iff (***) \quad \mathbb{P}\left(\max_{z \in \mathcal{Z}} \left| \sum_{i=1}^{d} Y_i(z) - \mathbb{E}\, Y_i(z) \right| \geqslant d\varepsilon \right) \leqslant \delta.$$

Union-bound:

$$\mathbb{P}\left(\max_{z \in \mathcal{Z}} |\dots| \geqslant d\varepsilon \right) \leqslant |\mathcal{Z}| \max_{z \in \mathcal{Z}} \mathbb{P}\left(|\dots| \geqslant d\varepsilon \right)$$

This term can be bounded independently of z using concentration bounds for sub-gamma RVs (bottom of page 6).

$$\mathbb{P}\left(|\dots| \geqslant d\varepsilon \right) \leqslant 2 \exp\left\{ -\frac{1}{4}\left(\frac{d\varepsilon}{4e\sigma^2} \wedge \frac{d^2\varepsilon^2}{d(4e\sigma^2)^2} \right) \right\}$$

absolute value

Take $\varepsilon = 1/2$ in the theorem page 6, and using $t = d\varepsilon$.

In addition, $|\mathcal{Z}| \leqslant n^2$.

$$\Rightarrow \mathbb{P}\left(\max_{z \in \mathcal{Z}} |\dots| \geqslant d\varepsilon \right) \leqslant \underbrace{2n^2 \exp\left\{ -\frac{d\varepsilon}{16e\sigma^2}\left(1 \wedge \frac{\varepsilon}{4e\sigma^2}\right) \right\}}$$

We want this term to be less than $\delta \in (0,1)$.

$\Rightarrow$ Select d appropriately:

$$2n^2 \exp\left\{ -\frac{d\varepsilon}{16e\sigma^2}\left(1 \wedge \frac{\varepsilon}{4e\sigma^2}\right) \right\} \leqslant \delta$$

$$\iff -\frac{d\varepsilon}{16e\sigma^2}\left(1 \wedge \frac{\varepsilon}{4e\sigma^2}\right) \leqslant \log\left(\frac{\delta}{2n^2}\right)$$

$$\iff d\left(1 \wedge \frac{\varepsilon}{4e\sigma^2}\right) \geqslant \frac{16e\sigma^2}{\varepsilon} \log\left(\frac{2n^2}{\delta}\right)$$

$$\iff d \geqslant \frac{32e\sigma^2}{\varepsilon}\left(1 \vee \frac{4e\sigma^2}{\varepsilon}\right) \log\left(\sqrt{\frac{2}{\delta}}\, n\right),$$

which concludes the proof. ∎

III - APPLICATION TO K-MEANS.

Random projections are particularly well-suited as a pre-processing step for the K-means algorithm, since it preserves the distance between points.

K-means revisited =

- Input: $x_1, \dots, x_n \in \mathbb{R}^D$, T = linear transformation
- Compute: $y_1, \dots, y_n$; $y_i = T x_i \in \mathbb{R}^d$

  ↘ Select $c_1^0, \dots, c_K^0 \in \mathbb{R}^d$

  ↘ Construct clusters

  $$\mathcal{C}_j^0 = \{ y_i \mid \| y_i - c_j^0 \|_2 \leq \| y_i - c_\ell^0 \|_2 \quad \forall \ell \neq j \}$$

  ↘ Repeat $m = 0, 1, \dots$

  - $c_j^{m+1} := \frac{1}{|\mathcal{C}_j^m|} \sum_{i \in \mathcal{C}_j^m} y_i$
  - $\mathcal{C}_j^{m+1} := \{ y_i \mid \| y_i - c_j^{m+1} \|_2 \leq \| y_i - c_\ell^{m+1} \| \quad \forall \ell \neq j \}$

- Output: $\mathcal{C}_j^{final}$ / c_j^{final}.

  and cluster $x_1, \dots, x_n$ such that

  $$\mathcal{C}_j := \{ x_i \mid y_i = T x_i \in \mathcal{C}_j^{final} \} , \quad 1 \leq j \leq K$$

↑ Performance bounds in Biau, Devroye, Lugosi (2008)
On the performance of clustering in Hilbert Spaces.
IEEE Trans. on Information Theory, Vol 54, Issue 2