Distance Learning for high dimensional data

Pablo Groisman
with M. Jonckheere and F. Sapienza

University of Buenos Aires and IMAS-CONICET

Sin Ciencia, Tecnología e Innovación Productiva... No hay futuro.

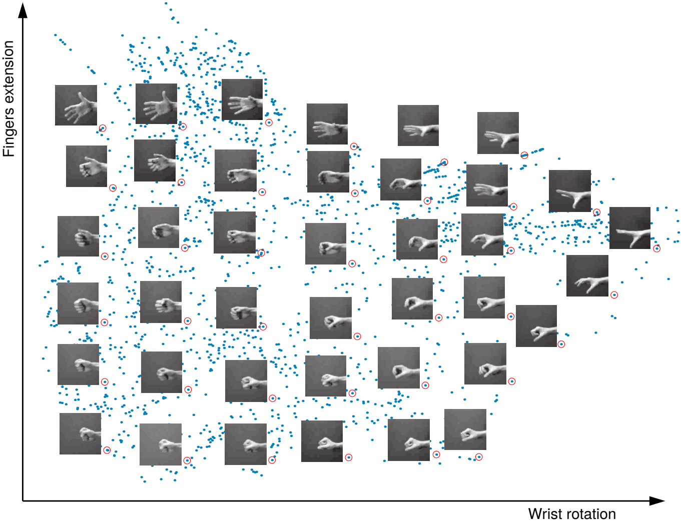
La magnitud de los probelmas...

...nos obliga a potenciar y promover la producción y transmisión del conocimiento, reconociendo a éste como el principal bien social y estratégico de las naciones para garantizar la mejora sostenible de la calidad de vida de sus habitantes.

Directorio Conicet

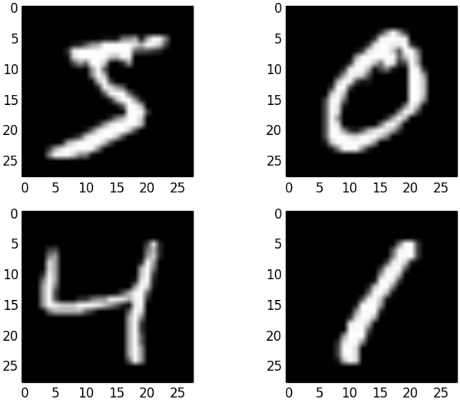
09-2018

Motivation: Hands Images



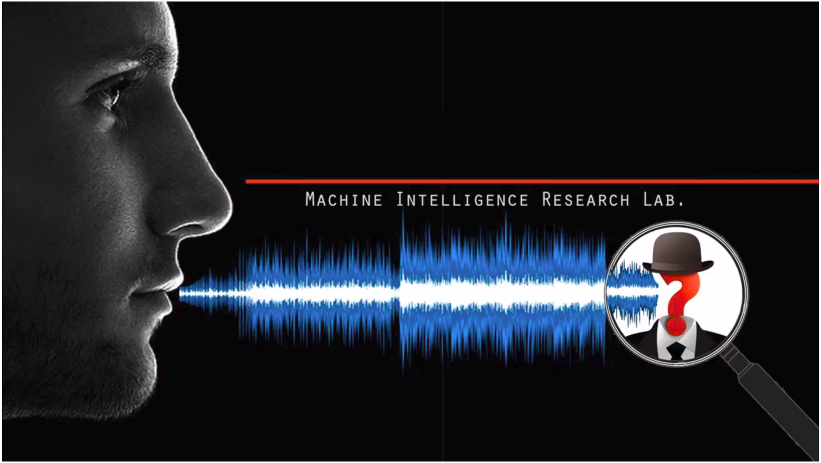
©J. B. Tenenbaum, V. de Silva, J. C. Langford, Science (2000)

Motivation: MNIST Dataset



©MNIST Dataset

Motivation: Speaker Identification



Motivation: A problem at Aristas SRL

Problem

- Clustering of high dimensional chemical formulas
- Distance between them in terms of e.g. olfactory properties

Data size

- 10^6 formulas
- Dimension $d \sim 4000$

A curse of dimensionality

A curse of dimensionality

Let $\omega_D(r) = \omega_D(1) r^D$ be the volume of the ball of radius r in $\mathbb{R}^D$.

$$\frac{\omega_D(1) - \omega_D(1-\varepsilon)}{\omega_D(1)} = 1 - (1-\varepsilon)^D \xrightarrow{D \to \infty} 1$$

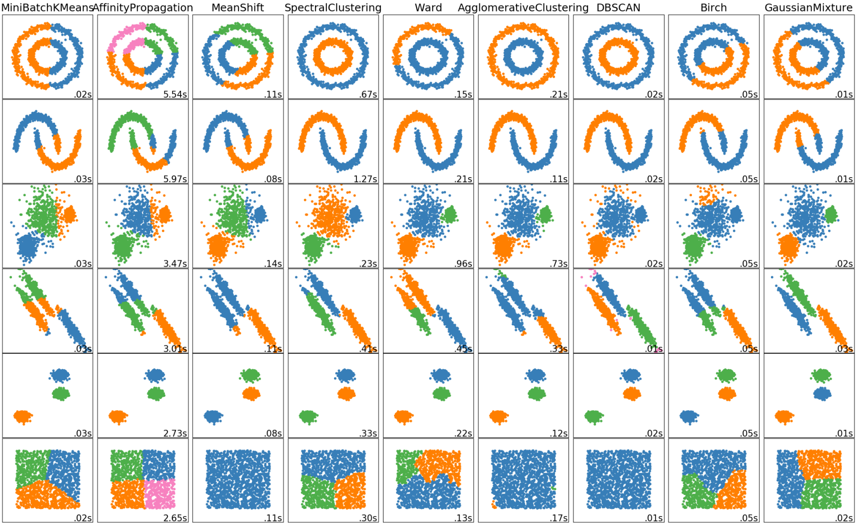
A curse of dimensionality

Let $\omega_D(r) = \omega_D(1) r^D$ be the volume of the ball of radius r in $\mathbb{R}^D$.

$$\frac{\omega_D(1) - \omega_D(1-\varepsilon)}{\omega_D(1)} = 1 - (1-\varepsilon)^D \xrightarrow{D \to \infty} 1$$

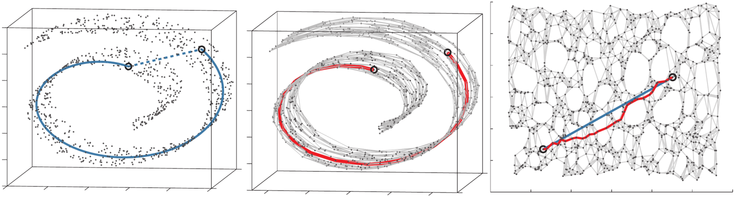
In high dimensional Euclidean spaces every two points of a typical large set are at similar distance.

Clustering: K-means, DBSCAN, etc.



Dimensionality Reduction: Isomap

Constructs the k-nn graph and finds the optimal path. The weight of an edge is given $|q_i - q_j|$.



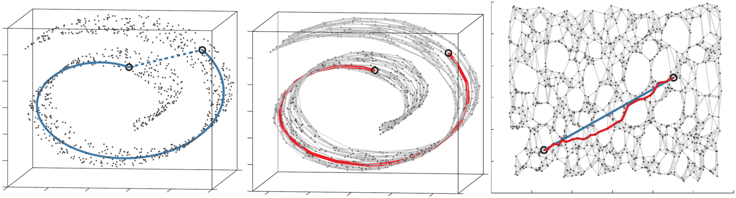
Dimensionality Reduction: Isomap

Theorem

Given $\varepsilon > 0$ and $\delta > 0$, for n large enough

$$\mathbb{P}\left(1-\varepsilon \leq \frac{d_{\mathit{geodesic}}(x,y)}{d_{\mathit{graph}}(x,y)} \leq 1+\varepsilon\right) > 1-\delta.$$

[Bernstein, de Silva, Langford, Tenenbaum (2000)].



Motivation

In most unsupervised learning tasks, a notion of similarity between data points is both crucial and usually not directly available as an input.

Motivation

In most unsupervised learning tasks, a notion of similarity between data points is both crucial and usually not directly available as an input.

The efficiency of tasks like dimensionality reduction and clustering might crucially depend on the distance chosen.

Motivation

In most unsupervised learning tasks, a notion of similarity between data points is both crucial and usually not directly available as an input.

The efficiency of tasks like dimensionality reduction and clustering might crucially depend on the distance chosen.

Since the data lies in an (unknown) lower dimensional surface, this distance has to be inferred from the data itself.

Motivation

In most unsupervised learning tasks, a notion of similarity between data points is both crucial and usually not directly available as an input.

The efficiency of tasks like dimensionality reduction and clustering might crucially depend on the distance chosen.

Since the data lies in an (unknown) lower dimensional surface, this distance has to be inferred from the data itself.

Motivation

We look for a distance that takes into account the underlying structure (surface) of the data and the underlying density from which the points are sampled.

The Problem

Let $\mathscr{M} \subseteq \mathbb{R}^D$ be a d-dimensional surface (we expect $d \ll D$).

The Problem

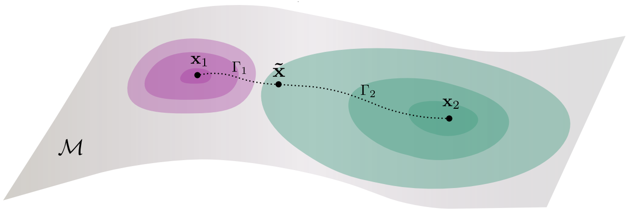
Let $\mathscr{M} \subseteq \mathbb{R}^D$ be a d-dimensional surface (we expect $d \ll D$).

Consider n independent points on $\mathscr{M}$ with common density $f : \mathscr{M} \mapsto \mathbb{R}_{\geq 0}$.

The Problem

Let $\mathscr{M} \subseteq \mathbb{R}^D$ be a d-dimensional surface (we expect $d \ll D$).

Consider n independent points on $\mathscr{M}$ with common density $f : \mathscr{M} \mapsto \mathbb{R}_{\geq 0}$.



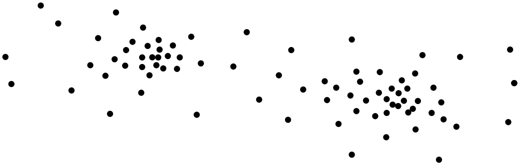
¿Can we learn the structure of $\mathscr{M}$?

Dimension, distances between points, etc.

The Problem

Let $\mathscr{M} \subseteq \mathbb{R}^D$ be a d-dimensional surface (we expect $d \ll D$).

Consider n independent points on $\mathscr{M}$ with common density $f : \mathscr{M} \mapsto \mathbb{R}_{\geq 0}$.



¿Can we learn the structure of $\mathscr{M}$?

Dimension, distances between points, etc.

Sample Fermat's distance

$\alpha \geq 1$ a parameter, $\mathbb{X} =$ a discrete set of points q, $x, y \in \mathbb{X}$

Sample Fermat's distance

$\alpha \geq 1$ a parameter, $\mathbb{X} =$ a discrete set of points q, $x, y \in \mathbb{X}$

$r_{xy} = (q_1, \ldots, q_K)$ an $\mathbb{X}$-path from x to y

$$\mathscr{T}(r_{xy}) = \sum_{j=1}^{K-1} |q_{i+1} - q_i|^{\alpha},$$

Sample Fermat's distance

$\alpha \geq 1$ a parameter, $\mathbb{X} =$ a discrete set of points q, $x, y \in \mathbb{X}$
$r_{xy} = (q_1, \ldots, q_K)$ an $\mathbb{X}$-path from x to y

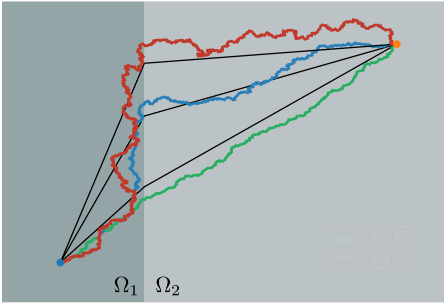
$$\mathscr{T}(r_{xy}) = \sum_{j=1}^{K-1} |q_{i+1} - q_i|^{\alpha}, \qquad D_{\mathbb{X}}(x, y) = \inf \mathscr{T}(r_{xy})$$

Sample Fermat's distance

$\alpha \geq 1$ a parameter, $\mathbb{X}$ = a discrete set of points q, $x, y \in \mathbb{X}$

$r_{xy} = (q_1, \ldots, q_K)$ an $\mathbb{X}$-path from x to y

$$\mathscr{T}(r_{xy}) = \sum_{j=1}^{K-1} |q_{i+1} - q_i|^\alpha, \qquad D_{\mathbb{X}}(x, y) = \inf \mathscr{T}(r_{xy})$$



The optimal path for $\alpha = 1.5$, $\alpha = 3$ and $\alpha = 6$.

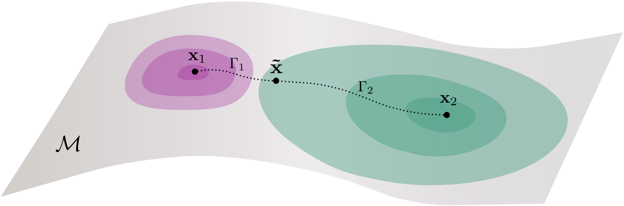
The density of points $\mathbb{X}$ in Ω_1 is higher than in Ω_2.

Fermat's Distance

$f : \mathscr{M} \to \mathbb{R}$ a density.

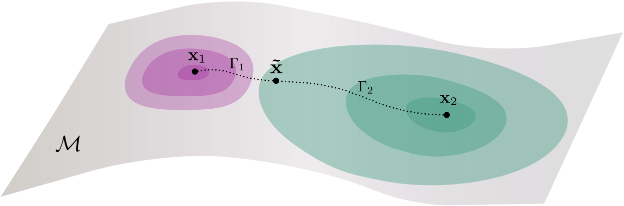
Fermat's Distance

$f : \mathcal{M} \to \mathbb{R}$ a density.



Fermat's Distance

$f : \mathscr{M} \to \mathbb{R}$ a density.



For $x, y \in \mathscr{M}$ and $\beta \geq 0$ we define **Fermat's distance** by

$$\mathscr{D}(x,y) = \inf_{\Gamma} \int_{\Gamma} \frac{1}{f^{\beta}} d\ell,$$

the minimization is over all curves Γ from x to y.

Fermat's principle

In optics, the path taken between two points by a ray of light is an extreme of the functional

$$\Gamma \mapsto \int_\Gamma \mathrm{n}(x) d\ell, \quad \mathrm{n} = \text{refractive index}$$

Fermat's principle

In optics, the path taken between two points by a ray of light is an extreme of the functional

$$\Gamma \mapsto \int_\Gamma \mathrm{n}(x) d\ell, \quad \mathrm{n} = \text{refractive index}$$

$$\mathscr{D}(x,y) = \inf_\Gamma \int_\Gamma \frac{1}{f^\beta} \qquad f^{-\beta} \sim \mathbf{n}$$

Fermat's principle

In optics, the path taken between two points by a ray of light is an extreme of the functional

$$\Gamma \mapsto \int_\Gamma \mathrm{n}(x) d\ell, \quad \mathrm{n} = \text{refractive index}$$

$$\mathscr{D}(x,y) = \inf_\Gamma \int_\Gamma \frac{1}{f^\beta} \qquad f^{-\beta} \sim \mathbf{n}$$

©S.Thorgerson - Pink Floyd, The Dark Side of the Moon (1973), Harvest, Capitol.

Fermat's distance

Theorem

For $x, y \in \mathscr{M}$ and $\mathbb{X}_n$ i.i.d $\sim f$ we have

$$\lim_{n\to\infty} n^{\beta} D_{\mathbb{X}_n}(x,y) = \mathscr{D}(x,y)$$

with $\beta = (\alpha - 1)/d$.

Fermat's distance

Theorem

For $x, y \in \mathscr{M}$ *and* $\mathbb{X}_n$ *i.i.d* $\sim f$ *we have*

$$\lim_{n \to \infty} n^{\beta} D_{\mathbb{X}_n}(x, y) = \mathscr{D}(x, y)$$

with $\beta = (\alpha - 1)/d$.

Heuristics:

$r = (q_1, \ldots, q_k)$ a path

$$\sum |q_{i+1} - q_i|^{\alpha} = \sum |q_{i+1} - q_i|^{\alpha - 1} |q_{i+1} - q_i|$$

Heuristics:

$r = (q_1, \ldots, q_k)$ a path

$$\sum |q_{i+1} - q_i|^\alpha = \sum |q_{i+1} - q_i|^{\alpha-1} |q_{i+1} - q_i|$$

Heuristics:

$r = (q_1, \ldots, q_k)$ a path

$$\sum |q_{i+1} - q_i|^\alpha = \sum |q_{i+1} - q_i|^{\alpha-1} |q_{i+1} - q_i|$$

$$nc_d |q_{i+1} - q_i|^d f(q_i) \asymp$$

Heuristics:

$r = (q_1, \ldots, q_k)$ a path

$$\sum |q_{i+1} - q_i|^\alpha = \sum |q_{i+1} - q_i|^{\alpha-1} |q_{i+1} - q_i|$$

$$nc_d |q_{i+1} - q_i|^d f(q_i) \asymp 1 \iff$$

Heuristics:

$r = (q_1, \dots, q_k)$ a path

$$\sum |q_{i+1} - q_i|^\alpha = \sum |q_{i+1} - q_i|^{\alpha-1} |q_{i+1} - q_i|$$

$$nc_d |q_{i+1} - q_i|^d f(q_i) \asymp 1 \iff n^{1/d} |q_{i+1} - q_i| \asymp c \frac{1}{f(q_i)^{1/d}}$$

Heuristics:

$r = (q_1, \ldots, q_k)$ a path

$$\sum |q_{i+1} - q_i|^\alpha = \sum |q_{i+1} - q_i|^{\alpha-1} |q_{i+1} - q_i|$$

$$n c_d |q_{i+1} - q_i|^d f(q_i) \asymp 1 \iff n^{1/d} |q_{i+1} - q_i| \asymp c \frac{1}{f(q_i)^{1/d}}$$

$$n^{(\alpha-1)/d} |q_{i+1} - q_i|^{\alpha-1} \asymp c \frac{1}{f(q_i)^{(\alpha-1)/d}}$$

Heuristics:

$r = (q_1, \ldots, q_k)$ a path

$$\sum |q_{i+1} - q_i|^\alpha = \sum |q_{i+1} - q_i|^{\alpha-1} |q_{i+1} - q_i|$$

$$n c_d |q_{i+1} - q_i|^d f(q_i) \asymp 1 \iff n^{1/d} |q_{i+1} - q_i| \asymp c \frac{1}{f(q_i)^{1/d}}$$

$$n^{(\alpha-1)/d} |q_{i+1} - q_i|^{\alpha-1} \asymp c \frac{1}{f(q_i)^{(\alpha-1)/d}}$$

$$\inf_r n^{(\alpha-1)/d} \sum |q_{i+1} - q_i|^\alpha \asymp \inf_\Gamma \int_\Gamma \frac{1}{f^\beta} \, d\ell.$$

Heuristics:

$r = (q_1, \ldots, q_k)$ a path

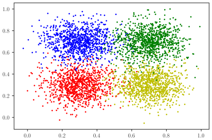
$$\sum |q_{i+1} - q_i|^\alpha = \sum |q_{i+1} - q_i|^{\alpha-1} |q_{i+1} - q_i|$$

$$n c_d |q_{i+1} - q_i|^d f(q_i) \asymp 1 \iff n^{1/d} |q_{i+1} - q_i| \asymp c \frac{1}{f(q_i)^{1/d}}$$

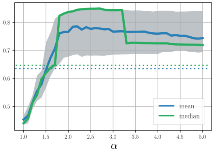
$$n^{(\alpha-1)/d} |q_{i+1} - q_i|^{\alpha-1} \asymp c \frac{1}{f(q_i)^{(\alpha-1)/d}}$$

$$\inf_r n^{(\alpha-1)/d} \sum |q_{i+1} - q_i|^\alpha \asymp \inf_\Gamma \int_\Gamma \frac{1}{f^\beta} \, d\ell. \qquad \square$$

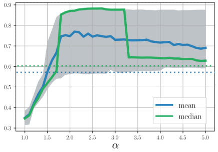
Clustering with Fermat K-medoids in the Swiss roll



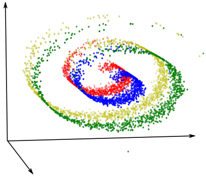
(a) 2D data



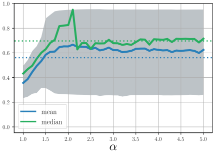
(c) Adjusted mutual information



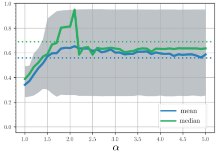
(e) Adjusted Rand index



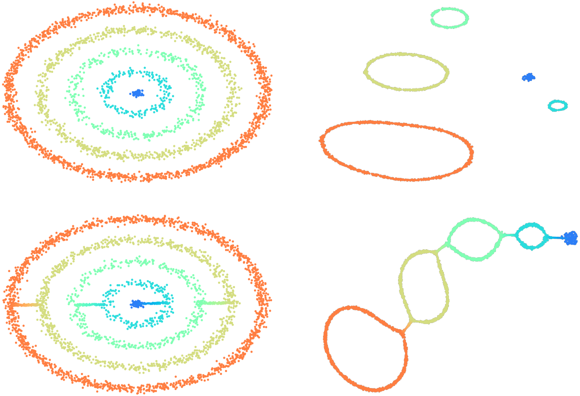
(b) 3D data



(d) Accuracy



(f) F1 score



Algorithmic considerations

Restricted Fermat's distance:

$$\mathbb{D}_{\mathbb{X}}^{(k)}(x,y) = \inf_{\substack{r=(q_1,\ldots,q_K) \\ q_{i+1}\in\mathscr{N}_k(q_i)}} \sum_{k=1}^{K-1} |q_{i+1}-q_i|^{\alpha}.$$

Algorithmic considerations

Restricted Fermat's distance:

$$\mathbb{D}_{\mathbb{X}}^{(k)}(x,y) = \inf_{\substack{r=(q_1,\ldots,q_K)\\ q_{i+1}\in\mathcal{N}_k(q_i)}} \sum_{k=1}^{K-1} |q_{i+1}-q_i|^{\alpha}.$$

Proposition: *Given $\varepsilon > 0$, we can choose $k = \mathcal{O}(\log(n/\varepsilon))$ such that*

$$\mathbb{P}\left(D_{\mathbb{X}_n}^{(k)}(x,y) = D_{\mathbb{X}_n}(x,y)\right) > 1-\varepsilon.$$

Algorithmic considerations

Restricted Fermat's distance:

$$\mathbb{D}_{\mathbb{X}}^{(k)}(x,y) = \inf_{\substack{r=(q_1,\dots,q_K)\\ q_{i+1}\in\mathcal{N}_k(q_i)}} \sum_{k=1}^{K-1} |q_{i+1}-q_i|^{\alpha}.$$

Proposition: *Given $\varepsilon > 0$, we can choose $k = \mathcal{O}(\log(n/\varepsilon))$ such that*

$$\mathbb{P}\left(D_{\mathbb{X}_n}^{(k)}(x,y) = D_{\mathbb{X}_n}(x,y)\right) > 1-\varepsilon.$$

$\rightarrow$ We can reduce the running time from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2(\log n)^2)$.

Conclusions

- We have introduced Fermat's distance and way to estimate it with a sample.

Conclusions

- We have introduced Fermat's distance and way to estimate it with a sample.
- It defines a notion of distance between sample points that takes into account the geometry of the clouds of point, including possible non-homogeneities.

Conclusions

- We have introduced Fermat's distance and way to estimate it with a sample.
- It defines a notion of distance between sample points that takes into account the geometry of the clouds of point, including possible non-homogeneities.
- We have proved that this estimator in fact approximates Fermat's distance, which is a good way to measure distance in this (general) setting.

Applications

- Clustering

Applications

- Clustering
- Dimensionality reduction

Applications

- Clustering
- Dimensionality reduction
- Density estimation

Applications

- Clustering
- Dimensionality reduction
- Density estimation
- Regression

Applications

- Clustering
- Dimensionality reduction
- Density estimation
- Regression
- Any learning task that requires a notion of distance as an input.

Applications

- Clustering
- Dimensionality reduction
- Density estimation
- Regression
- Any learning task that requires a notion of distance as an input.
- Any other task that requires a notion of distance as an input.

Applications

- Clustering
- Dimensionality reduction
- Density estimation
- Regression
- Any learning task that requires a notion of distance as an input.
- Any other task that requires a notion of distance as an input.
- etc.

Download

A prototype implementation is available at

`https://github.com/facusapienza21/Fermat_distance`

Scan me

Thanks!