Linear Projections of High-Dimensional Data

Elizabeth Meckes

Case Western Reserve University

LDHD Summer School
SAMSI
August, 2013

Why project?

Why project?

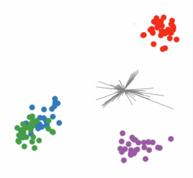
- To make expensive computations/algorithms feasible – so-called *Dimension reduction*

Why project?

- To make expensive computations/algorithms feasible – so-called *Dimension reduction*
- To visualize the data and look for global structure

Why project?

- To make expensive computations/algorithms feasible – so-called *Dimension reduction*
- To visualize the data and look for global structure



J. Faith

2-D projection of expression levels of 100 genes for samples from four tumor types

A quick word from our sponsor: Random Subspaces

A quick word from our sponsor: Random Subspaces

Definition: The Stiefel manifold $\mathfrak{W}_{d,k}$ is the set of ordered k-tuples of orthonormal vectors in $\mathbb{R}^d$:

$$\mathfrak{W}_{d,k} := \left\{ (v_1, \ldots, v_k) \in (\mathbb{R}^d)^k \,\middle|\, \langle v_i, v_j \rangle = \delta_{ij} \right\}.$$

A quick word from our sponsor: Random Subspaces

Definition: The Stiefel manifold $\mathfrak{W}_{d,k}$ is the set of ordered k-tuples of orthonormal vectors in $\mathbb{R}^d$:

$$\mathfrak{W}_{d,k} := \left\{ (v_1, \ldots, v_k) \in (\mathbb{R}^d)^k \,\middle|\, \langle v_i, v_j \rangle = \delta_{ij} \right\}.$$

How to pick a random element of $\mathfrak{W}_{d,k}$:

- Pick v_1 uniformly from $\mathbb{S}^{d-1}$.
- Pick v_2 uniformly from the unit sphere in $v_1^\perp$.
- Continue in the obvious way.

A quick word from our sponsor: Random Subspaces

Definition: The Stiefel manifold $\mathfrak{W}_{d,k}$ is the set of ordered k-tuples of orthonormal vectors in $\mathbb{R}^d$:

$$\mathfrak{W}_{d,k} := \left\{ (v_1, \ldots, v_k) \in (\mathbb{R}^d)^k \,\middle|\, \langle v_i, v_j \rangle = \delta_{ij} \right\}.$$

How to pick a random element of $\mathfrak{W}_{d,k}$:

- Pick v_1 uniformly from $\mathbb{S}^{d-1}$.
- Pick v_2 uniformly from the unit sphere in $v_1^\perp$.
- Continue in the obvious way.

The probability measure (called Haar measure) constructed this way is the unique rotation-invariant probability on $\mathfrak{W}_{d,k}$: if $U \in \mathbb{O}(d)$ is fixed, then

$$(v_1, \ldots, v_k) \stackrel{\mathcal{L}}{=} (Uv_1, \ldots, Uv_k).$$

Concentration of measure on $\mathfrak{W}_{d,k}$

$\mathfrak{W}_{d,k}$ is a metric space: if $\theta = (\theta_1, \ldots, \theta_k)$ and $\theta' = (\theta'_1, \ldots, \theta'_k)$, then we define the distance $\rho(\theta, \theta')$ between them by

$$\rho(\theta, \theta') := \sqrt{\sum_{i=1}^{k} \|\theta_i - \theta'_i\|^2}.$$

Concentration of measure on $\mathfrak{W}_{d,k}$

$\mathfrak{W}_{d,k}$ is a metric space: if $\theta = (\theta_1, \ldots, \theta_k)$ and $\theta' = (\theta'_1, \ldots, \theta'_k)$, then we define the distance $\rho(\theta, \theta')$ between them by

$$\rho(\theta, \theta') := \sqrt{\sum_{i=1}^{k} \|\theta_i - \theta'_i\|^2}.$$

Theorem (Milman–Schechtman)

There are constants C, c (independent of d and k) such that if $F : \mathfrak{W}_{d,k} \to \mathbb{R}$ is Lipschitz with Lipschitz constant L and Θ is a random point of $\mathfrak{W}_{d,k}$, then

$$\mathbb{P}\Big[\big|F(\Theta) - \mathbb{E}F(\Theta)\big| > L\epsilon\Big] \le Ce^{-cd\epsilon^2}.$$

The Johnson–Lindenstrauss Lemma

The Johnson–Lindenstrauss Lemma

If you have n high-dimensional data points and project them onto a random subspace of dimension $\sim \log(n)$, the pairwise distances between the points is approximately preserved.

The Johnson–Lindenstrauss Lemma

If you have n high-dimensional data points and project them onto a random subspace of dimension $\sim \log(n)$, the pairwise distances between the points is approximately preserved.

Practical conclusion: If your problem is about the metric structure of the data (finding the closest pair, most separated pair, minimum spanning tree of a graph,etc.), there is no need to work in the high-dimensional space that the data naturally live in.

The Johnson–Lindenstrauss Lemma

Lemma (J–L)

Let $\{x_j\}_{j=1}^{n} \subseteq \mathbb{R}^{d}$, and let U be a random $k \times d$ matrix, constructed by taking $U = V^T$ where the columns of V are the entries of a random point of $\mathfrak{W}_{d,k}$; that is,

U is a projection of $\mathbb{R}^d$ onto a random k-dimensional subspace.

The Johnson–Lindenstrauss Lemma

Lemma (J–L)

Let $\{x_j\}_{j=1}^{n} \subseteq \mathbb{R}^d$, and let U be a random $k \times d$ matrix, constructed by taking $U = V^T$ where the columns of V are the entries of a random point of $\mathfrak{W}_{d,k}$; that is,

U is a projection of $\mathbb{R}^d$ onto a random k-dimensional subspace.

If $k = \frac{a \log(n)}{\epsilon^2}$, then with probability $1 - \frac{C}{n^{\frac{ac}{9}-2}}$ (with C, c coming from the concentration inequality),

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \left(\frac{d}{k}\right)\|Ux_i - Ux_j\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2$$

for all $i, j \in \{1, \ldots, n\}$.

Application: Finding the closest point

Application: Finding the closest point

Consider the following problem: You are given a reference set $\mathcal{X}$ of n points in $\mathbb{R}^d$. Now given a query point $q \in \mathbb{R}^d$, find the closest point in $\mathcal{X}$ to q.

P. Indyk

dimension = number of pixels

Application: Finding the closest point

Consider the following problem: You are given a reference set $\mathcal{X}$ of n points in $\mathbb{R}^d$. Now given a query point $q \in \mathbb{R}^d$, find the closest point in $\mathcal{X}$ to q.

P. Indyk

dimension = number of pixels

The naïve approach – calculate each distance and keep track of the best so far – runs in $O(nd)$ steps.

Application: Finding the closest point

Surely you can relax a little

$\sim$ let Bill and Joram help you out.

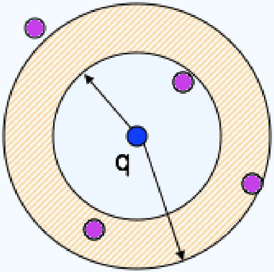
Application: Finding the closest point

Surely you can relax a little

$\sim$ let Bill and Joram help you out.

If you project onto a random subspace of dimension about $\log(n)$, distances are approximately preserved.

This means that while the algorithm might not return the absolute closest point, the point that it returns will be almost as close to q as the true closest point is.



P. Indyk

More carefully, suppose that U is one of the good random projections so that

$$(1-\epsilon)\|q-x_i\|^2 \leq \left(\frac{d}{k}\right)\|Uq-Ux_i\|^2 \leq (1+\epsilon)\|q-x_i\|^2$$

for each i.

More carefully, suppose that U is one of the good random projections so that

$$(1-\epsilon)\|q-x_i\|^2 \leq \left(\frac{d}{k}\right)\|Uq-Ux_i\|^2 \leq (1+\epsilon)\|q-x_i\|^2$$

for each i.

If Ux_i is the closest point to Uq (and so our randomized algorithm returns x_i), but the true closest point to q is x_j, then

$$\|q-x_i\| \leq (1+\epsilon)\|q-x_j\|;$$

that is, the wrong answer isn't *that* wrong.

More carefully, suppose that U is one of the good random projections so that

$$(1-\epsilon)\|q-x_i\|^2 \leq \left(\frac{d}{k}\right)\|Uq-Ux_i\|^2 \leq (1+\epsilon)\|q-x_i\|^2$$

for each i.

If Ux_i is the closest point to Uq (and so our randomized algorithm returns x_i), but the true closest point to q is x_j, then

$$\|q-x_i\| \leq (1+\epsilon)\|q-x_j\|;$$

that is, the wrong answer isn't *that* wrong.

And after projecting, the naïve approach runs in $O(n\log(n))$ steps, instead of $O(n^2)$.

Proof

We want to show that for each pair (i, j),

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \left(\frac{d}{k}\right)\left\|Ux_i - Ux_j\right\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2$$

with high probability,

Proof

We want to show that for each pair (i, j),

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \left(\frac{d}{k}\right)\|Ux_i - Ux_j\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2$$

with high probability, or equivalently,

$$\sqrt{1-\epsilon} \leq \sqrt{\frac{d}{k}}\|Ux\| \leq \sqrt{1+\epsilon}$$

for $x := \frac{x_i - x_j}{\|x_i - x_j\|}$.

Proof

We want to show that for each pair (i, j),

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \left(\frac{d}{k}\right) \left\|Ux_i - Ux_j\right\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2$$

with high probability, or equivalently,

$$\sqrt{1-\epsilon} \leq \sqrt{\frac{d}{k}}\|Ux\| \leq \sqrt{1+\epsilon}$$

for $x := \frac{x_i - x_j}{\|x_i - x_j\|}$.

By construction of U, this is the same as

$$\sqrt{1-\epsilon} \leq \sqrt{\frac{d}{k}}\left\|\left(\langle\theta_1, x\rangle, \ldots, \langle\theta_k, x\rangle\right)\right\| \leq \sqrt{1+\epsilon},$$

where $\theta = (\theta_1, \ldots, \theta_k)$ is a random point of $\mathfrak{W}_{d,k}$.

Proof, ctd.

For $x \in \mathbb{S}^{d-1}$ fixed, consider the function $F_x : \mathfrak{W}_{d,k} \to \mathbb{R}$ defined by

$$F_x(\theta_1, \ldots, \theta_k) = \sqrt{\frac{d}{k}} \left\| \left(\langle \theta_1, x \rangle, \ldots, \langle \theta_k, x \rangle \right) \right\|.$$

Proof, ctd.

For $x \in \mathbb{S}^{d-1}$ fixed, consider the function $F_x : \mathfrak{W}_{d,k} \to \mathbb{R}$ defined by

$$F_x(\theta_1, \ldots, \theta_k) = \sqrt{\frac{d}{k}} \left\| \left(\langle \theta_1, x \rangle, \ldots, \langle \theta_k, x \rangle \right) \right\|.$$

Now, if $\theta, \theta' \in \mathfrak{W}_{d,k}$, then

$$\left| \left\| \left(\langle \theta_1, x \rangle, \ldots, \langle \theta_k, x \rangle \right) \right\| - \left\| \left(\langle \theta'_1, x \rangle, \ldots, \langle \theta'_k, x \rangle \right) \right\| \right|$$

Proof, ctd.

For $x \in \mathbb{S}^{d-1}$ fixed, consider the function $F_x : \mathfrak{W}_{d,k} \to \mathbb{R}$ defined by

$$F_x(\theta_1, \ldots, \theta_k) = \sqrt{\frac{d}{k}} \left\| \left(\langle \theta_1, x \rangle, \ldots, \langle \theta_k, x \rangle \right) \right\|.$$

Now, if $\theta, \theta' \in \mathfrak{W}_{d,k}$, then

$$\left| \left\| \left(\langle \theta_1, x \rangle, \ldots, \langle \theta_k, x \rangle \right) \right\| - \left\| \left(\langle \theta'_1, x \rangle, \ldots, \langle \theta'_k, x \rangle \right) \right\| \right|$$
$$\leq \left\| \left(\langle \theta_1 - \theta'_1, x \rangle, \ldots, \langle \theta_k - \theta'_k, x \rangle \right) \right\|$$

Proof, ctd.

For $x \in \mathbb{S}^{d-1}$ fixed, consider the function $F_x : \mathfrak{W}_{d,k} \to \mathbb{R}$ defined by

$$F_x(\theta_1, \ldots, \theta_k) = \sqrt{\frac{d}{k}} \left\| \left(\langle \theta_1, x \rangle, \ldots, \langle \theta_k, x \rangle \right) \right\|.$$

Now, if $\theta, \theta' \in \mathfrak{W}_{d,k}$, then

$$\begin{aligned} &\left| \left\| \left(\langle \theta_1, x \rangle, \ldots, \langle \theta_k, x \rangle \right) \right\| - \left\| \left(\langle \theta'_1, x \rangle, \ldots, \langle \theta'_k, x \rangle \right) \right\| \right| \\ &\qquad \leq \left\| \left(\langle \theta_1 - \theta'_1, x \rangle, \ldots, \langle \theta_k - \theta'_k, x \rangle \right) \right\| \\ &\qquad = \sqrt{\sum_{j=1}^{k} \left\langle \theta_j - \theta'_j, x \right\rangle^2} \end{aligned}$$

Proof, ctd.

For $x \in \mathbb{S}^{d-1}$ fixed, consider the function $F_x : \mathfrak{W}_{d,k} \to \mathbb{R}$ defined by

$$F_x(\theta_1, \ldots, \theta_k) = \sqrt{\frac{d}{k}} \Big\| \big(\langle \theta_1, x \rangle, \ldots, \langle \theta_k, x \rangle \big) \Big\|.$$

Now, if $\theta, \theta' \in \mathfrak{W}_{d,k}$, then

$$\begin{aligned}
&\Big| \big\| \big(\langle \theta_1, x \rangle, \ldots, \langle \theta_k, x \rangle \big) \big\| - \big\| \big(\langle \theta_1', x \rangle, \ldots, \langle \theta_k', x \rangle \big) \big\| \Big| \\
&\qquad \le \Big\| \big(\langle \theta_1 - \theta_1', x \rangle, \ldots, \langle \theta_k - \theta_k', x \rangle \big) \Big\| \\
&\qquad\qquad = \sqrt{\sum_{j=1}^{k} \big\langle \theta_j - \theta_j', x \big\rangle^2} \le \sqrt{\sum_{j=1}^{k} \|\theta_j - \theta_j'\|^2}
\end{aligned}$$

Proof, ctd.

For $x \in \mathbb{S}^{d-1}$ fixed, consider the function $F_x : \mathfrak{W}_{d,k} \to \mathbb{R}$ defined by

$$F_x(\theta_1, \ldots, \theta_k) = \sqrt{\frac{d}{k}} \left\| \left(\langle \theta_1, x \rangle, \ldots, \langle \theta_k, x \rangle \right) \right\|.$$

Now, if $\theta, \theta' \in \mathfrak{W}_{d,k}$, then

$$\begin{aligned} &\left| \left\| \left(\langle \theta_1, x \rangle, \ldots, \langle \theta_k, x \rangle \right) \right\| - \left\| \left(\langle \theta'_1, x \rangle, \ldots, \langle \theta'_k, x \rangle \right) \right\| \right| \\ &\qquad \leq \left\| \left(\langle \theta_1 - \theta'_1, x \rangle, \ldots, \langle \theta_k - \theta'_k, x \rangle \right) \right\| \\ &\qquad\qquad = \sqrt{\sum_{j=1}^{k} \left\langle \theta_j - \theta'_j, x \right\rangle^2} \leq \sqrt{\sum_{j=1}^{k} \|\theta_j - \theta'_j\|^2} = \rho(\theta, \theta'). \end{aligned}$$

Proof, ctd.

That is, the function

$$F_x(\theta_1, \ldots, \theta_k) = \sqrt{\frac{d}{k}} \left\| \left(\langle \theta_1, x \rangle, \ldots, \langle \theta_k, x \rangle \right) \right\|$$

is $\sqrt{\frac{d}{k}}$-Lipschitz on $\mathfrak{W}_{d,k}$.

Proof, ctd.

That is, the function

$$F_x(\theta_1, \ldots, \theta_k) = \sqrt{\frac{d}{k}} \left\| \left(\langle \theta_1, x \rangle, \ldots, \langle \theta_k, x \rangle \right) \right\|$$

is $\sqrt{\frac{d}{k}}$-Lipschitz on $\mathfrak{W}_{d,k}$.

It follows immediately from concentration of measure that

$$\mathbb{P}\left[|F_x(\theta) - \mathbb{E}F_x(\theta)| \geq \epsilon \right] \leq Ce^{-ck\epsilon^2}.$$

Proof, ctd.

That is, the function

$$F_x(\theta_1, \ldots, \theta_k) = \sqrt{\frac{d}{k}} \left\| \left(\langle \theta_1, x \rangle, \ldots, \langle \theta_k, x \rangle \right) \right\|$$

is $\sqrt{\frac{d}{k}}$-Lipschitz on $\mathfrak{W}_{d,k}$.

It follows immediately from concentration of measure that

$$\mathbb{P}\left[|F_x(\theta) - \mathbb{E}F_x(\theta)| \geq \epsilon \right] \leq Ce^{-ck\epsilon^2}.$$

Remember that $k = \frac{a \log(n)}{\epsilon^2}$, so we have that

$$\mathbb{P}\left[|F_x(\theta) - \mathbb{E}F_x(\theta)| \geq \epsilon \right] \leq \frac{C}{n^{ac}}.$$

Proof, ctd.

We need that $\mathbb{E}F_x(\theta) \approx 1$.

Proof, ctd.

We need that $\mathbb{E}F_x(\theta) \approx 1$.

By the invariance of Haar measure under translation and transposition,

$$\big(\langle\theta_1, x\rangle, \ldots, \langle\theta_k, x\rangle\big) \stackrel{\mathcal{L}}{=} \big(\langle\theta_1, e_1\rangle, \ldots, \langle\theta_k, e_1\rangle\big)$$

Proof, ctd.

We need that $\mathbb{E}F_x(\theta) \approx 1$.

By the invariance of Haar measure under translation and transposition,

$$\big(\langle\theta_1, x\rangle, \ldots, \langle\theta_k, x\rangle\big) \stackrel{\mathcal{L}}{=} \big(\langle\theta_1, e_1\rangle, \ldots, \langle\theta_k, e_1\rangle\big) \stackrel{\mathcal{L}}{=} (v_1, \ldots, v_k),$$

where v is distributed uniformly on $\mathbb{S}^{d-1} \subseteq \mathbb{R}^d$.

Proof, ctd.

We need that $\mathbb{E} F_x(\theta) \approx 1$.

By the invariance of Haar measure under translation and transposition,

$$\left(\langle \theta_1, x \rangle, \ldots, \langle \theta_k, x \rangle \right) \stackrel{\mathcal{L}}{=} \left(\langle \theta_1, e_1 \rangle, \ldots, \langle \theta_k, e_1 \rangle \right) \stackrel{\mathcal{L}}{=} (v_1, \ldots, v_k),$$

where v is distributed uniformly on $\mathbb{S}^{d-1} \subseteq \mathbb{R}^d$.

That is,

$$F_x(\theta) \stackrel{\mathcal{L}}{=} \sqrt{\left(\frac{d}{k}\right)\left(v_1^2 + \cdots + v_k^2\right)}.$$

Proof, ctd.

We need that $\mathbb{E}F_x(\theta) \approx 1$.

By the invariance of Haar measure under translation and transposition,

$$\left(\langle \theta_1, x \rangle, \ldots, \langle \theta_k, x \rangle \right) \stackrel{\mathcal{L}}{=} \left(\langle \theta_1, e_1 \rangle, \ldots, \langle \theta_k, e_1 \rangle \right) \stackrel{\mathcal{L}}{=} (v_1, \ldots, v_k),$$

where v is distributed uniformly on $\mathbb{S}^{d-1} \subseteq \mathbb{R}^d$.

That is,

$$F_x(\theta) \stackrel{\mathcal{L}}{=} \sqrt{\left(\frac{d}{k}\right)(v_1^2 + \cdots + v_k^2)}.$$

It is an easy excercise that $\mathbb{E}v_i^2 = \frac{1}{d}$ (so $\mathbb{E}\left[F_x(\theta)\right]^2 = 1$) and the concentration we already have for $F_x(\theta)$ then implies that $\mathbb{E}F_x(\theta) \approx 1$.

Proof, ctd.

So: returning to the original formulation, we have that

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \left(\frac{d}{k}\right)\left\|Ux_i - Ux_j\right\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2$$

with probability at least $1 - \frac{C}{n^{\frac{ac}{9}}}$.

Proof, ctd.

So: returning to the original formulation, we have that

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \left(\frac{d}{k}\right) \left\|Ux_i - Ux_j\right\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2$$

with probability at least $1 - \frac{C}{n^{\frac{ac}{9}}}$.

There are fewer than n^2 pairs (i, j), so a simple union bound gives that the above statement holds *for all pairs* (i, j) with probability at least $1 - \frac{C}{n^{\frac{ac}{9}-2}}$.

The Diaconis–Freedman Effect

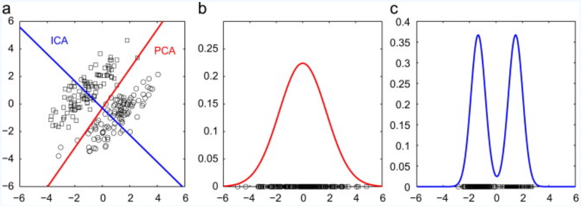
The Diaconis–Freedman Effect

If you project a large, high-dimensional data set onto one or two dimensions, what you get nearly always looks Gaussian, no matter what structure you started with.

The Diaconis–Freedman Effect

If you project a large, high-dimensional data set onto one or two dimensions, what you get nearly always looks Gaussian, no matter what structure you started with.

Practical conclusion: When looking for projections that tell you something interesting about the data, look for something that is very different from Gaussian.



The Diaconis–Freedman Effect

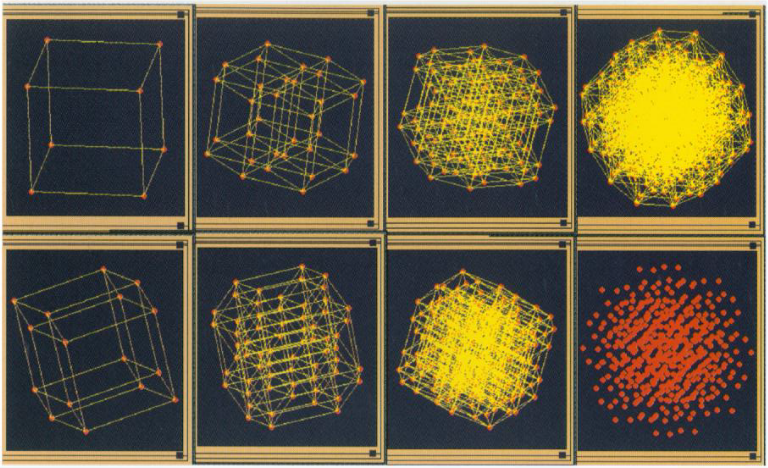


Figure from Buja, Cook, and Swayne "Interactive High-dimensional Data Visualization", 1996.

The Diaconis–Freedman Effect

Many authors have proved rigorous results that capture the D-F effect; e.g.,

- Sudakov (1978)
- Diaconis–Freedman (1984)
- von Weiszäcker (1997)
- Bobkov (2003)
- Klartag (2007)
- Dümbgen–Zerial (2011)
- ...

We'll focus on:

Theorem (E.M.)

Let $\{x_j\}_{j=1}^{n}$ *be data points in* $\mathbb{R}^d$

We'll focus on:

Theorem (E.M.)

Let $\{x_j\}_{j=1}^n$ *be data points in* $\mathbb{R}^d$*, satisfying*

- $\frac{1}{n}\sum_{j=1}^n x_j = \mathbf{0}$*, and* $\frac{1}{n}\sum_{j=1}^n |x_j|^2 = \sigma^2 d$*,*

We'll focus on:

Theorem (E.M.)

Let $\{x_j\}_{j=1}^n$ be data points in $\mathbb{R}^d$, satisfying

- $\frac{1}{n}\sum_{j=1}^n x_j = \mathbf{0}$, *and* $\frac{1}{n}\sum_{j=1}^n |x_j|^2 = \sigma^2 d$,
- $\sup_{\xi \in \mathbb{S}^{d-1}} \frac{1}{n}\sum_{j=1}^n \langle \xi, x_j \rangle^2 \leq L'$

We'll focus on:

Theorem (E.M.)

Let $\{x_j\}_{j=1}^n$ be data points in $\mathbb{R}^d$, satisfying

- $\frac{1}{n}\sum_{j=1}^n x_j = \mathbf{0}$, *and* $\frac{1}{n}\sum_{j=1}^n |x_j|^2 = \sigma^2 d$,
- $\sup_{\xi\in\mathbb{S}^{d-1}} \frac{1}{n}\sum_{j=1}^n \langle \xi, x_j\rangle^2 \leq L'$
- $\frac{1}{n}\sum_{j=1}^n \left|\frac{|x_j|^2}{d} - \sigma^2\right| \leq \frac{L}{\sqrt{d}}$.

We'll focus on:

Theorem (E.M.)

Let $\{x_j\}_{j=1}^n$ be data points in $\mathbb{R}^d$, satisfying

- $\frac{1}{n}\sum_{j=1}^n x_j = \mathbf{0}$, *and* $\frac{1}{n}\sum_{j=1}^n |x_j|^2 = \sigma^2 d$,
- $\sup_{\xi \in \mathbb{S}^{d-1}} \frac{1}{n}\sum_{j=1}^n \langle \xi, x_j \rangle^2 \leq L'$
- $\frac{1}{n}\sum_{j=1}^n \left| \frac{|x_j|^2}{d} - \sigma^2 \right| \leq \frac{L}{\sqrt{d}}$.

Let $E \subseteq \mathbb{R}^d$ be a random k-dimensional subspace and let μ_E denote the empirical measure of the projection of the $\{x_j\}$ onto E.

We'll focus on:

Theorem (E.M.)

Let $\{x_j\}_{j=1}^n$ be data points in $\mathbb{R}^d$, satisfying

- $\frac{1}{n}\sum_{j=1}^n x_j = \mathbf{0}$, *and* $\frac{1}{n}\sum_{j=1}^n |x_j|^2 = \sigma^2 d$,
- $\sup_{\xi \in \mathbb{S}^{d-1}} \frac{1}{n}\sum_{j=1}^n \langle \xi, x_j \rangle^2 \leq L'$
- $\frac{1}{n}\sum_{j=1}^n \left| \frac{|x_j|^2}{d} - \sigma^2 \right| \leq \frac{L}{\sqrt{d}}$.

Let $E \subseteq \mathbb{R}^d$ be a random k-dimensional subspace and let μ_E denote the empirical measure of the projection of the $\{x_j\}$ onto E. Then

$$\mathbb{E} d_{BL}(\mu_E, \sigma Z) \leq C \frac{k + \log(d)}{k^{\frac{2}{3}} d^{\frac{2}{3k+4}}} \tag{1}$$

We'll focus on:

Theorem (E.M.)

Let $\{x_j\}_{j=1}^n$ be data points in $\mathbb{R}^d$, satisfying

- $\frac{1}{n}\sum_{j=1}^n x_j = \mathbf{0}$, *and* $\frac{1}{n}\sum_{j=1}^n |x_j|^2 = \sigma^2 d$,
- $\sup_{\xi \in \mathbb{S}^{d-1}} \frac{1}{n}\sum_{j=1}^n \left\langle \xi, x_j \right\rangle^2 \leq L'$
- $\frac{1}{n}\sum_{j=1}^n \left| \frac{|x_j|^2}{d} - \sigma^2 \right| \leq \frac{L}{\sqrt{d}}$.

Let $E \subseteq \mathbb{R}^d$ be a random k-dimensional subspace and let μ_E denote the empirical measure of the projection of the $\{x_j\}$ onto E. Then

(1) $$\mathbb{E} d_{BL}(\mu_E, \sigma Z) \leq C \frac{k + \log(d)}{k^{\frac{2}{3}} d^{\frac{2}{3k+4}}}$$

(2) $$\mathbb{P}\left[\left| d_{BL}(\mu_E, \sigma Z) - \mathbb{E} d_{BL}(\mu_E, \sigma Z) \right| > \epsilon \right] \leq C e^{-cd\epsilon^2}.$$

Preliminaries to the proof

Let X be distributed uniformly in $\{x_1, \ldots, x_n\}$;

i.e., X is a randomly chosen data point.

Preliminaries to the proof

Let X be distributed uniformly in $\{x_1, \ldots, x_n\}$;

i.e., X is a randomly chosen data point.

For a k-dimensional subspace $E \subseteq \mathbb{R}^d$,

let X_E be distributed uniformly in $\{\pi_E(x_1,), \ldots, \pi_E(x_n)\}$;

i.e., X_E is the projection of X onto the subspace E.

Preliminaries to the proof

Let X be distributed uniformly in $\{x_1, \ldots, x_n\}$;

i.e., X is a randomly chosen data point.

For a k-dimensional subspace $E \subseteq \mathbb{R}^d$,

let X_E be distributed uniformly in $\{\pi_E(x_1,), \ldots, \pi_E(x_n)\}$;

i.e., X_E is the projection of X onto the subspace E.

There are two ways we might like to understand X_E:

1. "Annealed" behavior: X and E are both random and independent.
2. "Quenched" behavior: X is random but E is fixed; what is "typical"?

Outline of the proof

Step 1: The annealed projection X_E, when both X and E are random and independent, is approximately Gaussian.

This is done via Stein's method.

Outline of the proof

Step 1: The annealed projection X_E, when both X and E are random and independent, is approximately Gaussian.

This is done via Stein's method.

Step 2: The average distance to average $\mathbb{E}\left[d_{BL}(X_E, X_F)\right]$, where E is random inside the distance, but F is averaged over after measuring the distance, is small.

The bounded-Lipschitz distance is interpreted as the supremum of a stochastic process indexed by a class of test functions. Concentration of measure and entropy methods can then be used to derive a bound.

Outline of the proof

Step 1: The annealed projection X_E, when both X and E are random and independent, is approximately Gaussian.

This is done via Stein's method.

Step 2: The average distance to average $\mathbb{E}\left[d_{BL}(X_E, X_F)\right]$, where E is random inside the distance, but F is averaged over after measuring the distance, is small.

The bounded-Lipschitz distance is interpreted as the supremum of a stochastic process indexed by a class of test functions. Concentration of measure and entropy methods can then be used to derive a bound.

Step 3: The (random) bounded-Lipschitz distance $d_{BL}(X_E, X_F)$ is tightly concentrated near its mean.

This also follows from concentration of measure.

Outline of the proof – Stiefel manifold formulation

Step 1: The annealed projection X_Θ, when both X and Θ are random and independent, is approximately Gaussian.

This is done via Stein's method.

Step 2: The mean bounded-Lipschitz distance $\mathbb{E}_\theta d_{BL}(X_\theta, X_\Theta)$ is small.

The bounded-Lipschitz distance is interpreted as the supremum of a stochastic process indexed by a class of test functions. Concentration of measure and entropy methods can then be used to derive a bound.

Step 3: The (random) bounded-Lipschitz distance $d_{BL}(X_\theta, X_\Theta)$ is tightly concentrated near its mean.

This also follows from concentration of measure.

Outline of the proof – Stiefel manifold formulation

Step 3: The (random) bounded-Lipschitz distance $d_{BL}(X_\theta, X_\Theta)$ is tightly concentrated near its mean.

This also follows from concentration of measure.

More about Step 3

Consider the function $F : \mathfrak{W}_{d,k} \longrightarrow \mathbb{R}$ defined by

$$F(\theta) := d_{BL}(X_\theta, Y) = \sup_{\substack{|f| \leq 1, \\ f \text{ 1-Lipschitz}}} \left| \mathbb{E}f(X_\theta) - \mathbb{E}f(Y) \right|,$$

where Y is any reference distribution.

More about Step 3

Consider the function $F : \mathfrak{W}_{d,k} \to \mathbb{R}$ defined by

$$F(\theta) := d_{BL}(X_\theta, Y) = \sup_{\substack{|f| \le 1, \\ f \text{ 1-Lipschitz}}} \left| \mathbb{E}f(X_\theta) - \mathbb{E}f(Y) \right|,$$

where Y is any reference distribution. Then

$$\left| \left| \mathbb{E}f(X_\theta) - \mathbb{E}f(Y) \right| - \left| \mathbb{E}f(X_{\theta'}) - \mathbb{E}f(Y) \right| \right|$$

More about Step 3

Consider the function $F : \mathfrak{W}_{d,k} \to \mathbb{R}$ defined by

$$F(\theta) := d_{BL}(X_\theta, Y) = \sup_{\substack{|f| \leq 1, \\ f \text{ 1-Lipschitz}}} \left| \mathbb{E} f(X_\theta) - \mathbb{E} f(Y) \right|,$$

where Y is any reference distribution. Then

$$\begin{aligned} &\left| \left| \mathbb{E} f(X_\theta) - \mathbb{E} f(Y) \right| - \left| \mathbb{E} f(X_{\theta'}) - \mathbb{E} f(Y) \right| \right| \\ &\qquad \leq \left| \mathbb{E} f\left(\langle X, \theta_1 \rangle, \ldots, \langle X, \theta_k \rangle \right) - \mathbb{E} f\left(\langle X, \theta_1' \rangle, \ldots, \langle X, \theta_k' \rangle \right) \right| \end{aligned}$$

More about Step 3

Consider the function $F : \mathfrak{W}_{d,k} \to \mathbb{R}$ defined by

$$F(\theta) := d_{BL}(X_\theta, Y) = \sup_{\substack{|f| \le 1, \\ f \text{ 1-Lipschitz}}} \left| \mathbb{E} f(X_\theta) - \mathbb{E} f(Y) \right|,$$

where Y is any reference distribution. Then

$$\begin{aligned} &\Big| \left| \mathbb{E} f(X_\theta) - \mathbb{E} f(Y) \right| - \left| \mathbb{E} f(X_{\theta'}) - \mathbb{E} f(Y) \right| \Big| \\ &\qquad \le \left| \mathbb{E} f\left(\langle X, \theta_1 \rangle, \ldots, \langle X, \theta_k \rangle \right) - \mathbb{E} f\left(\langle X, \theta_1' \rangle, \ldots, \langle X, \theta_k' \rangle \right) \right| \\ &\qquad \le \mathbb{E} \left| \left(\langle X, \theta_1 - \theta_1' \rangle, \ldots, \langle X, \theta_k - \theta_k' \rangle \right) \right| \end{aligned}$$

More about Step 3

Consider the function $F : \mathfrak{W}_{d,k} \to \mathbb{R}$ defined by

$$F(\theta) := d_{BL}(X_\theta, Y) = \sup_{\substack{|f| \leq 1, \\ f \text{ 1-Lipschitz}}} \left| \mathbb{E} f(X_\theta) - \mathbb{E} f(Y) \right|,$$

where Y is any reference distribution. Then

$$\begin{aligned}
&\left| \left| \mathbb{E} f(X_\theta) - \mathbb{E} f(Y) \right| - \left| \mathbb{E} f(X_{\theta'}) - \mathbb{E} f(Y) \right| \right| \\
&\qquad \leq \left| \mathbb{E} f\left(\langle X, \theta_1 \rangle, \ldots, \langle X, \theta_k \rangle \right) - \mathbb{E} f\left(\langle X, \theta_1' \rangle, \ldots, \langle X, \theta_k' \rangle \right) \right| \\
&\qquad \leq \mathbb{E} \left| \left(\langle X, \theta_1 - \theta_1' \rangle, \ldots, \langle X, \theta_k - \theta_k' \rangle \right) \right| \\
&\qquad\qquad \leq \sqrt{\sum_{j=1}^{k} |\theta_j - \theta_j'|^2 \, \mathbb{E} \left\langle X, \frac{\theta_j - \theta_j'}{|\theta_j - \theta_j'|} \right\rangle^2}
\end{aligned}$$

More about Step 3

Consider the function $F : \mathfrak{W}_{d,k} \to \mathbb{R}$ defined by

$$F(\theta) := d_{BL}(X_\theta, Y) = \sup_{\substack{|f| \leq 1, \\ f \text{ 1-Lipschitz}}} \left| \mathbb{E} f(X_\theta) - \mathbb{E} f(Y) \right|,$$

where Y is any reference distribution. Then

$$\begin{aligned}
&\Big| \left| \mathbb{E} f(X_\theta) - \mathbb{E} f(Y) \right| - \left| \mathbb{E} f(X_{\theta'}) - \mathbb{E} f(Y) \right| \Big| \\
&\qquad \leq \left| \mathbb{E} f\left(\langle X, \theta_1 \rangle, \dots, \langle X, \theta_k \rangle \right) - \mathbb{E} f\left(\langle X, \theta_1' \rangle, \dots, \langle X, \theta_k' \rangle \right) \right| \\
&\qquad \leq \mathbb{E} \left| \left(\langle X, \theta_1 - \theta_1' \rangle, \dots, \langle X, \theta_k - \theta_k' \rangle \right) \right| \\
&\qquad\qquad \leq \sqrt{\sum_{j=1}^{k} |\theta_j - \theta_j'|^2 \, \mathbb{E} \left\langle X, \frac{\theta_j - \theta_j'}{|\theta_j - \theta_j'|} \right\rangle^2} \leq \rho(\theta, \theta') \sqrt{L'}
\end{aligned}$$

That is,

$$\Big| \big|\mathbb{E}f(X_\theta) - \mathbb{E}f(Y)\big| - \big|\mathbb{E}f(X_{\theta'}) - \mathbb{E}f(Y)\big| \Big| \leq \rho(\theta, \theta')\sqrt{L'},$$

That is,

$$\Big|\big|\mathbb{E}f(X_\theta) - \mathbb{E}f(Y)\big| - \big|\mathbb{E}f(X_{\theta'}) - \mathbb{E}f(Y)\big|\Big| \le \rho(\theta,\theta')\sqrt{L'},$$

so that

$$\begin{aligned} &\Big|F(\theta) - F(\theta')\Big| \\ &\qquad = \Big|\sup_f \big|\mathbb{E}f(X_\theta) - \mathbb{E}f(Y)\big| - \sup_f \big|\mathbb{E}f(X_{\theta'}) - \mathbb{E}f(Y)\big|\Big| \end{aligned}$$

That is,

$$\left| \left|\mathbb{E}f(X_\theta) - \mathbb{E}f(Y)\right| - \left|\mathbb{E}f(X_{\theta'}) - \mathbb{E}f(Y)\right| \right| \leq \rho(\theta, \theta')\sqrt{L'},$$

so that

$$\begin{aligned}
&\left|F(\theta) - F(\theta')\right| \\
&\qquad = \left| \sup_f \left|\mathbb{E}f(X_\theta) - \mathbb{E}f(Y)\right| - \sup_f \left|\mathbb{E}f(X_{\theta'}) - \mathbb{E}f(Y)\right| \right| \\
&\qquad \leq \sup_f \left| \left|\mathbb{E}f(X_\theta) - \mathbb{E}f(Y)\right| - \left|\mathbb{E}f(X_{\theta'}) - \mathbb{E}f(Y)\right| \right|
\end{aligned}$$

That is,

$$\left| \left|\mathbb{E}f(X_\theta) - \mathbb{E}f(Y)\right| - \left|\mathbb{E}f(X_{\theta'}) - \mathbb{E}f(Y)\right| \right| \leq \rho(\theta, \theta')\sqrt{L'},$$

so that

$$\begin{aligned}
&\left|F(\theta) - F(\theta')\right| \\
&\qquad = \left| \sup_f \left|\mathbb{E}f(X_\theta) - \mathbb{E}f(Y)\right| - \sup_f \left|\mathbb{E}f(X_{\theta'}) - \mathbb{E}f(Y)\right| \right| \\
&\qquad \leq \sup_f \left| \left|\mathbb{E}f(X_\theta) - \mathbb{E}f(Y)\right| - \left|\mathbb{E}f(X_{\theta'}) - \mathbb{E}f(Y)\right| \right| \\
&\qquad \leq \rho(\theta, \theta')\sqrt{L'};
\end{aligned}$$

i.e., $F(\theta) = d_{BL}(X_\theta, Y)$ is a $\sqrt{L'}$-Lipschitz function of $\theta \in \mathfrak{W}_{d,k}$.

Since $d_{BL}(X_\theta, X_\Theta)$ is $\sqrt{L'}$-Lipschitz, concentration of measure on $\mathfrak{W}_{d,k}$ immediately yields

$$\mathbb{P}_\theta\left[\left|d_{BL}(X_\theta, X_\Theta) - \mathbb{E}d_{BL}(X_\theta, X_\Theta)\right| > \epsilon\right] \leq Ce^{\frac{cd\epsilon^2}{L'}}.$$

Since $d_{BL}(X_\theta, X_\Theta)$ is $\sqrt{L'}$-Lipschitz, concentration of measure on $\mathfrak{W}_{d,k}$ immediately yields

$$\mathbb{P}_\theta\left[\left|d_{BL}(X_\theta, X_\Theta) - \mathbb{E}d_{BL}(X_\theta, X_\Theta)\right| > \epsilon\right] \leq Ce^{\frac{cd\epsilon^2}{L'}}.$$

That is, the random distance $d_{BL}(X_\theta, X_\Theta)$ is usually within about $\frac{1}{\sqrt{d}}$ of the "average distance to average" $\mathbb{E}d_{BL}(X_\theta, X_\Theta)$.

Outline of the proof – Stiefel manifold formulation

Step 1: The annealed projection X_Θ, when both X and Θ are random and independent, is approximately Gaussian.

This is done via Stein's method.

Step 2: The mean bounded-Lipschitz distance $\mathbb{E}_\theta d_{BL}(X_\theta, X_\Theta)$ is small.

The bounded-Lipschitz distance is interpreted as the supremum of a stochastic process indexed by a class of test functions. Concentration of measure and entropy methods can then be used to derive a bound.

Step 3: The (random) bounded-Lipschitz distance $d_{BL}(X_\theta, X_\Theta)$ is tightly concentrated near its mean.

This also follows from concentration of measure.

Outline of the proof – Stiefel manifold formulation

Step 2: The mean bounded-Lipschitz distance $\mathbb{E}_\theta d_{BL}(X_\theta, X_\Theta)$ is small.

The bounded-Lipschitz distance is interpreted as the supremum of a stochastic process indexed by a class of test functions. Concentration of measure and entropy methods can then be used to derive a bound.

Step 2 – Average distance to average

Step 2 – Average distance to average

We need to estimate

$$\mathbb{E}_\theta d_{BL}(X_\theta, X_\Theta) = \mathbb{E}\left(\sup_{\|f\|_{BL}\le 1}\left|\mathbb{E}\left[f(X_\theta)\middle|\theta\right] - \mathbb{E}f(X_\Theta)\right|\right).$$

Step 2 – Average distance to average

We need to estimate

$$\mathbb{E}_\theta d_{BL}(X_\theta, X_\Theta) = \mathbb{E}\left(\sup_{\|f\|_{BL}\le 1} \left| \mathbb{E}\left[f(X_\theta) \middle| \theta \right] - \mathbb{E} f(X_\Theta) \right| \right).$$

If the stochastic process $\{X_f\}_{\|f\|_{BL}\le 1}$ is defined by

$$X_f := \mathbb{E}\left[f(X_\theta) \middle| \theta \right] - \mathbb{E} f(X_\Theta),$$

then what we want is $\mathbb{E}\sup_{\|f\|_{BL}\le 1} X_f$.

Step 2 – Average distance to average

We need to estimate

$$\mathbb{E}_\theta d_{BL}(X_\theta, X_\Theta) = \mathbb{E}\left(\sup_{\|f\|_{BL}\le 1}\left|\mathbb{E}\left[f(X_\theta)\middle|\theta\right] - \mathbb{E}f(X_\Theta)\right|\right).$$

If the stochastic process $\{X_f\}_{\|f\|_{BL}\le 1}$ is defined by

$$X_f := \mathbb{E}\left[f(X_\theta)\middle|\theta\right] - \mathbb{E}f(X_\Theta),$$

then what we want is $\mathbb{E}\sup_{\|f\|_{BL}\le 1} X_f$.

Applying measure concentration this time to $F(\theta) := \mathbb{E}\left[(f-g)(X_\theta)\middle|\theta\right]$ shows that the process has the property:

$$\mathbb{P}\left[\left|X_f - X_g\right| > \epsilon\right] \le Ce^{-\frac{cd\epsilon^2}{\|f-g\|_{BL}^2}}.$$

Theorem (usually attributed to Dudley; probably actually due to Pisier)

If a stochastic process $\{X_t\}_{t\in T}$ *indexed by the metric space* (T, δ) *satisfies the a sub-Gaussian increment condition*

$$\mathbb{P}\left[\left|X_t - X_s\right| > \epsilon\right] \leq Ce^{-\frac{\epsilon^2}{2\delta^2(s,t)}} \qquad \forall \epsilon > 0,$$

then

$$\mathbb{E}\sup_{t\in T} X_t \leq C\int_0^\infty \sqrt{\log N(T,\delta,\epsilon)}d\epsilon,$$

where $N(T,\delta,\epsilon)$ *is the* ϵ*-covering number of* T *with respect to the distance* δ.

Theorem (usually attributed to Dudley; probably actually due to Pisier)

If a stochastic process $\{X_t\}_{t \in T}$ *indexed by the metric space* (T, δ) *satisfies the a sub-Gaussian increment condition*

$$\mathbb{P}\left[\left|X_t - X_s\right| > \epsilon\right] \leq Ce^{-\frac{\epsilon^2}{2\delta^2(s,t)}} \qquad \forall \epsilon > 0,$$

then

$$\mathbb{E} \sup_{t \in T} X_t \leq C \int_0^\infty \sqrt{\log N(T, \delta, \epsilon)} d\epsilon,$$

where $N(T, \delta, \epsilon)$ *is the* ϵ*-covering number of* T *with respect to the distance* δ*.*

Recall that our process satisfies

$$\mathbb{P}\Big[\left|X_f - X_g\right| > \epsilon\Big] \leq Ce^{-\frac{cd\epsilon^2}{\|f-g\|_{BL}^2}}.$$

The question, then, is: if $BL_1^k := \left\{ f : \mathbb{R}^k \to \mathbb{R} \middle| \|f\|_{BL} \leq 1 \right\}$, what is $N\left(BL_1^k, \frac{\|\cdot\|_{BL}}{\sqrt{d}}, \epsilon\right)$?

The question, then, is: if $BL_1^k := \left\{ f : \mathbb{R}^k \to \mathbb{R} \middle| \|f\|_{BL} \leq 1 \right\}$, what is $N\left(BL_1^k, \frac{\|\cdot\|_{BL}}{\sqrt{d}}, \epsilon\right)$?

Bad news: $N\left(BL_1^k, \frac{\|\cdot\|_{BL}}{\sqrt{d}}, \epsilon\right) = \infty$.

The question, then, is: if $BL_1^k := \left\{ f : \mathbb{R}^k \to \mathbb{R} \middle| \|f\|_{BL} \le 1 \right\}$, what is $N\left(BL_1^k, \frac{\|\cdot\|_{BL}}{\sqrt{d}}, \epsilon\right)$?

Bad news: $N\left(BL_1^k, \frac{\|\cdot\|_{BL}}{\sqrt{d}}, \epsilon\right) = \infty$.

But not to worry: approximating Lipschitz functions by piecewise affine functions and using volumetric estimates in the resulting finite-dimensional normed space of approximating functions does the job, and ultimately we get

$$\mathbb{E}_\theta d_{BL}(X_\theta, X_\Theta) \le C \frac{k + \log(d)}{k^{\frac{2}{3}} d^{\frac{2}{3k+4}}}.$$