When is normal normal?

Quantitative asymptotics of graphical projection pursuit

Elizabeth Meckes

Case Western Reserve University

Cleveland State University, November 13, 2009

A limit result for projections of high-dimensional data

A limit result for projections of high-dimensional data

Theorem (Diaconis-Freedman)

A limit result for projections of high-dimensional data

Theorem (Diaconis-Freedman)

Let $x_1, \ldots, x_n$ be deterministic vectors in $\mathbb{R}^d$. Suppose that the x_i, n and d depend on a hidden index ν, such that as ν tends to infinity, so do n and d.

A limit result for projections of high-dimensional data

Theorem (Diaconis-Freedman)

Let $x_1, \ldots, x_n$ be deterministic vectors in $\mathbb{R}^d$. Suppose that the x_i, n and d depend on a hidden index ν, such that as ν tends to infinity, so do n and d. Suppose there exists $\sigma > 0$ such that, for any fixed $\epsilon > 0$,

A limit result for projections of high-dimensional data

Theorem (Diaconis-Freedman)

Let $x_1, \ldots, x_n$ be deterministic vectors in $\mathbb{R}^d$. Suppose that the x_i, n and d depend on a hidden index ν, such that as ν tends to infinity, so do n and d. Suppose there exists $\sigma > 0$ such that, for any fixed $\epsilon > 0$,

$$\frac{1}{n}\left|\{j \leq n : \left||x_j|^2 - \sigma^2 d\right| > \epsilon d\}\right| \xrightarrow{\nu \to \infty} 0 \tag{1}$$

A limit result for projections of high-dimensional data

Theorem (Diaconis-Freedman)

Let $x_1, \ldots, x_n$ be deterministic vectors in $\mathbb{R}^d$. Suppose that the x_i, n and d depend on a hidden index ν, such that as ν tends to infinity, so do n and d. Suppose there exists $\sigma > 0$ such that, for any fixed $\epsilon > 0$,

$$\frac{1}{n}\left|\left\{j \leq n : \left||x_j|^2 - \sigma^2 d\right| > \epsilon d\right\}\right| \xrightarrow{\nu \to \infty} 0 \tag{1}$$

and

$$\frac{1}{n^2}\left|\left\{j, k \leq n : \left|\left\langle x_j, x_k \right\rangle\right| > \epsilon d\right\}\right| \xrightarrow{\nu \to \infty} 0. \tag{2}$$

A limit result for projections of high-dimensional data

Theorem (Diaconis-Freedman)

Let $x_1, \ldots, x_n$ be deterministic vectors in $\mathbb{R}^d$. Suppose that the x_i, n and d depend on a hidden index ν, such that as ν tends to infinity, so do n and d. Suppose there exists $\sigma > 0$ such that, for any fixed $\epsilon > 0$,

$$\frac{1}{n}\left|\{j \leq n : \left||x_j|^2 - \sigma^2 d\right| > \epsilon d\}\right| \xrightarrow{\nu \to \infty} 0 \tag{1}$$

and

$$\frac{1}{n^2}\left|\{j,k \leq n : \left|\left\langle x_j, x_k \right\rangle\right| > \epsilon d\}\right| \xrightarrow{\nu \to \infty} 0. \tag{2}$$

Let θ be distributed according to normalized surface area measure on the unit sphere $\mathbb{S}^{d-1} \subseteq \mathbb{R}^d$ and let μ_ν^θ be the (random) probability measure on $\mathbb{R}$ which puts equal mass at each of the points $\{\langle \theta, x_i \rangle\}_{i=1}^n$.

A limit result for projections of high-dimensional data

Theorem (Diaconis-Freedman)

Let $x_1, \ldots, x_n$ be deterministic vectors in $\mathbb{R}^d$. Suppose that the x_i, n and d depend on a hidden index ν, such that as ν tends to infinity, so do n and d. Suppose there exists $\sigma > 0$ such that, for any fixed $\epsilon > 0$,

$$\frac{1}{n}\left|\left\{j \leq n : \left||x_j|^2 - \sigma^2 d\right| > \epsilon d\right\}\right| \xrightarrow{\nu \to \infty} 0 \tag{1}$$

and

$$\frac{1}{n^2}\left|\left\{j,k \leq n : \left|\left\langle x_j, x_k \right\rangle\right| > \epsilon d\right\}\right| \xrightarrow{\nu \to \infty} 0. \tag{2}$$

Let θ be distributed according to normalized surface area measure on the unit sphere $\mathbb{S}^{d-1} \subseteq \mathbb{R}^d$ and let μ_ν^θ be the (random) probability measure on $\mathbb{R}$ which puts equal mass at each of the points $\{\langle \theta, x_i \rangle\}_{i=1}^n$. Then, as $\nu \to \infty$, the measures μ_ν^θ tend to $\mathcal{N}(0, \sigma^2)$ weakly in probability.

Interpretation:

Interpretation: In a set of high-dimensional data in which the data vectors are all about the same length and don't concentrate too much in any one direction, most one-dimensional projections of the data will look about Gaussian.

Interpretation: In a set of high-dimensional data in which the data vectors are all about the same length and don't concentrate too much in any one direction, most one-dimensional projections of the data will look about Gaussian.

Possible conclusion:

Interpretation: In a set of high-dimensional data in which the data vectors are all about the same length and don't concentrate too much in any one direction, most one-dimensional projections of the data will look about Gaussian.

Possible conclusion: When trying to understand high-dimensional data through projections, look for projections which are not close to Gaussian.

Interpretation: In a set of high-dimensional data in which the data vectors are all about the same length and don't concentrate too much in any one direction, most one-dimensional projections of the data will look about Gaussian.

Possible conclusion: When trying to understand high-dimensional data through projections, look for projections which are not close to Gaussian.

Question 1:

Interpretation: In a set of high-dimensional data in which the data vectors are all about the same length and don't concentrate too much in any one direction, most one-dimensional projections of the data will look about Gaussian.

Possible conclusion: When trying to understand high-dimensional data through projections, look for projections which are not close to Gaussian.

Question 1: Is there a way to tell whether a projection that looks close to Gaussian is interesting?

Question 2: If the data are projected onto a k-dimensional subspace instead of a 1-dimensional subspace, does this phenomenon persist?

Interpretation: In a set of high-dimensional data in which the data vectors are all about the same length and don't concentrate too much in any one direction, most one-dimensional projections of the data will look about Gaussian.

Possible conclusion: When trying to understand high-dimensional data through projections, look for projections which are not close to Gaussian.

Question 1: Is there a way to tell whether a projection that looks close to Gaussian is interesting?

Question 2: If the data are projected onto a k-dimensional subspace instead of a 1-dimensional subspace, does this phenomenon persist? If so, how can k grow with n and d?

A possible answer to question 1:

A possible answer to question 1:

A projection which looks close to Gaussian may be interesting if it is “very close” to Gaussian.

A possible answer to question 1:

A projection which looks close to Gaussian may be interesting if it is “very close” to Gaussian. That is, if one could quantify how close to Gaussian (in some metric) a typical projection of data with no structure should be, then a projection which is much closer than that might be interesting.

A possible answer to question 1:

A projection which looks close to Gaussian may be interesting if it is “very close” to Gaussian. That is, if one could quantify how close to Gaussian (in some metric) a typical projection of data with no structure should be, then a projection which is much closer than that might be interesting.

$\Longrightarrow$ What is a good quantitative version of the theorem of Diaconis and Freedman?

A possible answer to question 1:

A projection which looks close to Gaussian may be interesting if it is “very close” to Gaussian. That is, if one could quantify how close to Gaussian (in some metric) a typical projection of data with no structure should be, then a projection which is much closer than that might be interesting.

$\Longrightarrow$ What is a good quantitative version of the theorem of Diaconis and Freedman?

Finding such a quantitative result is a possible route to an answer to question 2 as well. If a bound in terms of k can be found on the typical distance to Gaussian of k-dimensional projections, then some conclusions about the rate that k may grow with n and d are possible.

A weaker theorem

A weaker theorem

Let $x_1, \ldots, x_n$ be deterministic vectors in $\mathbb{R}^d$.

A weaker theorem

Let $x_1, \ldots, x_n$ be deterministic vectors in $\mathbb{R}^d$. Let θ be a random point on the sphere $\mathbb{S}^{d-1} \subseteq \mathbb{R}^d$ and let K be an independent random element of $\{1, \ldots, n\}$.

A weaker theorem

Let $x_1, \ldots, x_n$ be deterministic vectors in $\mathbb{R}^d$. Let θ be a random point on the sphere $\mathbb{S}^{d-1} \subseteq \mathbb{R}^d$ and let K be an independent random element of $\{1, \ldots, n\}$. Consider the random variable

$$W := \langle \theta, x_K \rangle .$$

A weaker theorem

Let $x_1, \ldots, x_n$ be deterministic vectors in $\mathbb{R}^d$. Let θ be a random point on the sphere $\mathbb{S}^{d-1} \subseteq \mathbb{R}^d$ and let K be an independent random element of $\{1, \ldots, n\}$. Consider the random variable

$$W := \langle \theta, x_K \rangle .$$

The distribution of W puts equal mass at each of the random points $\langle \theta, x_i \rangle$.

A weaker theorem

Let $x_1, \ldots, x_n$ be deterministic vectors in $\mathbb{R}^d$. Let θ be a random point on the sphere $\mathbb{S}^{d-1} \subseteq \mathbb{R}^d$ and let K be an independent random element of $\{1, \ldots, n\}$. Consider the random variable

$$W := \langle \theta, x_K \rangle .$$

The distribution of W puts equal mass at each of the random points $\langle \theta, x_i \rangle$. We'll show that W is approximately a Gaussian random variable, with an explicit bound on the total variation distance to Gaussian with d and n fixed.

What's so weak?

What's so weak?

This is a weak theorem because it concerns the distribution of the random variable $W = \langle \theta, x_K \rangle$, which has two sources of randomness:

What's so weak?

This is a weak theorem because it concerns the distribution of the random variable $W = \langle \theta, x_K \rangle$, which has two sources of randomness: K is random, which gives us an empirical distribution for the n data points;

What's so weak?

This is a weak theorem because it concerns the distribution of the random variable $W = \langle \theta, x_K \rangle$, which has two sources of randomness: K is random, which gives us an empirical distribution for the n data points; θ is also random, which means that instead of looking at a fixed (randomly chosen) projection of the data, we're looking at an averaged version of the empirical distribution over all possible projections.

What's so weak?

This is a weak theorem because it concerns the distribution of the random variable $W = \langle \theta, x_K \rangle$, which has two sources of randomness: K is random, which gives us an empirical distribution for the n data points; θ is also random, which means that instead of looking at a fixed (randomly chosen) projection of the data, we're looking at an averaged version of the empirical distribution over all possible projections.

The randomness in θ will be exploited in an essential way to prove that W is close to Gaussian.

Assumptions for the data

Assumptions for the data

Recall that the theorem of Diaconis and Freedman had the following assumptions on the vectors x_i:

Assumptions for the data

Recall that the theorem of Diaconis and Freedman had the following assumptions on the vectors x_i: There exists $\sigma > 0$ such that, for any fixed $\epsilon > 0$,

Assumptions for the data

Recall that the theorem of Diaconis and Freedman had the following assumptions on the vectors x_i: There exists $\sigma > 0$ such that, for any fixed $\epsilon > 0$,

$$\frac{1}{n}\left|\{j \leq n : \left||x_j|^2 - \sigma^2 d\right| > \epsilon d\}\right| \xrightarrow{\nu \to \infty} 0$$

Assumptions for the data

Recall that the theorem of Diaconis and Freedman had the following assumptions on the vectors x_i: There exists $\sigma > 0$ such that, for any fixed $\epsilon > 0$,

$$\frac{1}{n}\left|\{j \leq n : \left||x_j|^2 - \sigma^2 d\right| > \epsilon d\}\right| \xrightarrow{\nu \to \infty} 0$$

and

$$\frac{1}{n^2}\left|\{j, k \leq n : \left|\left\langle x_j, x_k \right\rangle\right| > \epsilon d\}\right| \xrightarrow{\nu \to \infty} 0.$$

Assumptions for the data

Recall that the theorem of Diaconis and Freedman had the following assumptions on the vectors x_i: There exists $\sigma > 0$ such that, for any fixed $\epsilon > 0$,

$$\frac{1}{n}\left|\{j \leq n : \left||x_j|^2 - \sigma^2 d\right| > \epsilon d\}\right| \xrightarrow{\nu\to\infty} 0$$

and

$$\frac{1}{n^2}\left|\{j,k \leq n : \left|\left\langle x_j, x_k\right\rangle\right| > \epsilon d\}\right| \xrightarrow{\nu\to\infty} 0.$$

We'll need quantitative versions of these limit conditions:

Assumptions for the data

Recall that the theorem of Diaconis and Freedman had the following assumptions on the vectors x_i: There exists $\sigma > 0$ such that, for any fixed $\epsilon > 0$,

$$\frac{1}{n}\left|\{j \leq n : \left||x_j|^2 - \sigma^2 d\right| > \epsilon d\}\right| \xrightarrow{\nu\to\infty} 0$$

and

$$\frac{1}{n^2}\left|\{j,k \leq n : \left|\left\langle x_j, x_k\right\rangle\right| > \epsilon d\}\right| \xrightarrow{\nu\to\infty} 0.$$

We'll need quantitative versions of these limit conditions: Suppose that σ is defined by the equation $\frac{1}{n}\sum \left|x_i\right|^2 = \sigma^2 d$ and that A and B are defined by

Assumptions for the data

Recall that the theorem of Diaconis and Freedman had the following assumptions on the vectors x_i: There exists $\sigma > 0$ such that, for any fixed $\epsilon > 0$,

$$\frac{1}{n}\left|\{j \leq n : \left||x_j|^2 - \sigma^2 d\right| > \epsilon d\}\right| \xrightarrow{\nu\to\infty} 0$$

and

$$\frac{1}{n^2}\left|\{j,k \leq n : \left|\left\langle x_j, x_k\right\rangle\right| > \epsilon d\}\right| \xrightarrow{\nu\to\infty} 0.$$

We'll need quantitative versions of these limit conditions: Suppose that σ is defined by the equation $\frac{1}{n}\sum \left|x_i\right|^2 = \sigma^2 d$ and that A and B are defined by

$$\frac{1}{n}\sum_{i=1}^{n}\left|\sigma^{-2}|x_i|^2 - d\right| =: A$$

Assumptions for the data

Recall that the theorem of Diaconis and Freedman had the following assumptions on the vectors x_i: There exists $\sigma > 0$ such that, for any fixed $\epsilon > 0$,

$$\frac{1}{n}\left|\{j \leq n : \left||x_j|^2 - \sigma^2 d\right| > \epsilon d\}\right| \xrightarrow{\nu\to\infty} 0$$

and

$$\frac{1}{n^2}\left|\{j,k \leq n : \left|\left\langle x_j, x_k\right\rangle\right| > \epsilon d\}\right| \xrightarrow{\nu\to\infty} 0.$$

We'll need quantitative versions of these limit conditions: Suppose that σ is defined by the equation $\frac{1}{n}\sum \left|x_i\right|^2 = \sigma^2 d$ and that A and B are defined by

$$\frac{1}{n}\sum_{i=1}^{n}\left|\sigma^{-2}|x_i|^2 - d\right| =: A$$

and

$$\sup_{\theta\in\mathbb{S}^{n-1}}\frac{1}{n}\sum_{i=1}^{n}\langle\theta, x_i\rangle^2 =: B.$$

An abstract normal approxiation theorem

An abstract normal approxiation theorem

Theorem (M)

Suppose that (W, W_ϵ) *is a family of exchangeable pairs defined on a common probability space, such that* $\mathbb{E}W = 0$ *and* $\mathbb{E}W^2 = \sigma^2$.

An abstract normal approxiation theorem

Theorem (M)

Suppose that (W, W_ϵ) *is a family of exchangeable pairs defined on a common probability space, such that* $\mathbb{E}W = 0$ *and* $\mathbb{E}W^2 = \sigma^2$. *Suppose there is a function* $\lambda(\epsilon)$ *and a random variable* E *such that*

An abstract normal approxiation theorem

Theorem (M)

Suppose that (W, W_ϵ) *is a family of exchangeable pairs defined on a common probability space, such that* $\mathbb{E}W = 0$ *and* $\mathbb{E}W^2 = \sigma^2$. *Suppose there is a function* $\lambda(\epsilon)$ *and a random variable E such that*

1. $\frac{1}{\lambda(\epsilon)}\mathbb{E}\left[W_\epsilon - W \middle| W\right] \xrightarrow[\epsilon \to 0]{L_1} -W.$

An abstract normal approxiation theorem

Theorem (M)

Suppose that (W, W_ϵ) *is a family of exchangeable pairs defined on a common probability space, such that* $\mathbb{E}W = 0$ *and* $\mathbb{E}W^2 = \sigma^2$. *Suppose there is a function* $\lambda(\epsilon)$ *and a random variable* E *such that*

1. $\frac{1}{\lambda(\epsilon)}\mathbb{E}\left[W_\epsilon - W \middle| W\right] \xrightarrow[\epsilon \to 0]{L_1} -W.$
2. $\frac{1}{2\lambda(\epsilon)\sigma^2}\mathbb{E}\left[(W_\epsilon - W)^2 \middle| W\right] \xrightarrow[\epsilon \to 0]{L_1} 1 + E.$

An abstract normal approxiation theorem

Theorem (M)

Suppose that (W, W_ϵ) *is a family of exchangeable pairs defined on a common probability space, such that* $\mathbb{E}W = 0$ *and* $\mathbb{E}W^2 = \sigma^2$. *Suppose there is a function* $\lambda(\epsilon)$ *and a random variable* E *such that*

1. $\frac{1}{\lambda(\epsilon)}\mathbb{E}\left[W_\epsilon - W \middle| W\right] \xrightarrow[\epsilon \to 0]{L_1} -W.$
2. $\frac{1}{2\lambda(\epsilon)\sigma^2}\mathbb{E}\left[(W_\epsilon - W)^2 \middle| W\right] \xrightarrow[\epsilon \to 0]{L_1} 1 + E.$
3. $\frac{1}{\lambda(\epsilon)}\mathbb{E}|W_\epsilon - W|^3 \xrightarrow{\epsilon \to 0} 0.$

An abstract normal approxiation theorem

Theorem (M)

Suppose that (W, W_ϵ) *is a family of exchangeable pairs defined on a common probability space, such that* $\mathbb{E}W = 0$ *and* $\mathbb{E}W^2 = \sigma^2$. *Suppose there is a function* $\lambda(\epsilon)$ *and a random variable* E *such that*

1. $\frac{1}{\lambda(\epsilon)}\mathbb{E}\left[W_\epsilon - W \middle| W\right] \xrightarrow[\epsilon \to 0]{L_1} -W.$
2. $\frac{1}{2\lambda(\epsilon)\sigma^2}\mathbb{E}\left[(W_\epsilon - W)^2 \middle| W\right] \xrightarrow[\epsilon \to 0]{L_1} 1 + E.$
3. $\frac{1}{\lambda(\epsilon)}\mathbb{E}|W_\epsilon - W|^3 \xrightarrow{\epsilon \to 0} 0.$

Then if Z *is a standard normal random variable,*

$$d_{TV}(W, Z) \leq \mathbb{E}\left|E\right|.$$

The exchangeable pair

The exchangeable pair

Let $A_\epsilon = \begin{bmatrix} \sqrt{1-\epsilon^2} & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2} \end{bmatrix} \oplus I_{n-2}$ and let $U \in \mathcal{O}_d$ be a random orthogonal matrix, independent of everything.

The exchangeable pair

Let $A_\epsilon = \begin{bmatrix} \sqrt{1-\epsilon^2} & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2} \end{bmatrix} \oplus I_{n-2}$ and let $U \in \mathcal{O}_d$ be a random orthogonal matrix, independent of everything. Let

$$W_\epsilon := \left\langle UA_\epsilon U^T \theta, x_K \right\rangle$$

The exchangeable pair

Let $A_\epsilon = \begin{bmatrix} \sqrt{1-\epsilon^2} & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2} \end{bmatrix} \oplus I_{n-2}$ and let $U \in \mathcal{O}_d$ be a random orthogonal matrix, independent of everything. Let

$$W_\epsilon := \left\langle UA_\epsilon U^T\theta, x_K \right\rangle \quad \Longrightarrow \quad W_\epsilon - W = \left\langle U(A_\epsilon - I_d)U^T\theta, x_K \right\rangle .$$

The exchangeable pair

Let $A_\epsilon = \begin{bmatrix} \sqrt{1-\epsilon^2} & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2} \end{bmatrix} \oplus I_{n-2}$ and let $U \in \mathcal{O}_d$ be a random orthogonal matrix, independent of everything. Let

$$W_\epsilon := \left\langle UA_\epsilon U^T\theta, x_K \right\rangle \quad \Longrightarrow \quad W_\epsilon - W = \left\langle U(A_\epsilon - I_d)U^T\theta, x_K \right\rangle.$$

Now, write $U(A_\epsilon - I_d)U^T = \epsilon HCH^T - \left(\frac{\epsilon^2}{2} + O(\epsilon^4)\right) HH^T$,

where H is the first two columns of U and $C = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$.

The exchangeable pair

Let $A_\epsilon = \begin{bmatrix} \sqrt{1-\epsilon^2} & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2} \end{bmatrix} \oplus I_{n-2}$ and let $U \in \mathcal{O}_d$ be a random orthogonal matrix, independent of everything. Let

$$W_\epsilon := \left\langle UA_\epsilon U^T\theta, x_K \right\rangle \quad \Longrightarrow \quad W_\epsilon - W = \left\langle U(A_\epsilon - I_d)U^T\theta, x_K \right\rangle.$$

Now, write $U(A_\epsilon - I_d)U^T = \epsilon HCH^T - \left(\frac{\epsilon^2}{2} + O(\epsilon^4)\right) HH^T$,

where H is the first two columns of U and $C = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$. It's easy to show that $\mathbb{E}\left[HCH^T\right] = 0$ and $\mathbb{E}\left[HH^T\right] = \frac{2}{d} I_d$,

The exchangeable pair

Let $A_\epsilon = \begin{bmatrix} \sqrt{1-\epsilon^2} & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2} \end{bmatrix} \oplus I_{n-2}$ and let $U \in \mathcal{O}_d$ be a random orthogonal matrix, independent of everything. Let

$$W_\epsilon := \left\langle UA_\epsilon U^T\theta, x_K \right\rangle \quad \Longrightarrow \quad W_\epsilon - W = \left\langle U(A_\epsilon - I_d)U^T\theta, x_K \right\rangle.$$

Now, write $U(A_\epsilon - I_d)U^T = \epsilon HCH^T - \left(\frac{\epsilon^2}{2} + O(\epsilon^4)\right) HH^T$, where H is the first two columns of U and $C = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$. It's easy to show that $\mathbb{E}\left[HCH^T\right] = 0$ and $\mathbb{E}\left[HH^T\right] = \frac{2}{d} I_d$, thus

$$\mathbb{E}\left[W_\epsilon - W \middle| W\right] = -\frac{\epsilon^2}{d} W + O(\epsilon^4).$$

The exchangeable pair

Let $A_\epsilon = \begin{bmatrix} \sqrt{1-\epsilon^2} & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2} \end{bmatrix} \oplus I_{n-2}$ and let $U \in \mathcal{O}_d$ be a random orthogonal matrix, independent of everything. Let

$$W_\epsilon := \left\langle UA_\epsilon U^T\theta, x_K \right\rangle \quad \Longrightarrow \quad W_\epsilon - W = \left\langle U(A_\epsilon - I_d)U^T\theta, x_K \right\rangle.$$

Now, write $U(A_\epsilon - I_d)U^T = \epsilon HCH^T - \left(\frac{\epsilon^2}{2} + O(\epsilon^4)\right) HH^T$,

where H is the first two columns of U and $C = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$. It's easy to show that $\mathbb{E}\left[HCH^T\right] = 0$ and $\mathbb{E}\left[HH^T\right] = \frac{2}{d} I_d$, thus

$$\mathbb{E}\left[W_\epsilon - W \middle| W\right] = -\frac{\epsilon^2}{d} W + O(\epsilon^4).$$

So $\lambda(\epsilon) = \frac{\epsilon^2}{d}$.

By the same decomposition and a bit more work,

By the same decomposition and a bit more work,

$$\mathbb{E}\left[(W_\epsilon - W)^2 \middle| W\right]$$
$$= \frac{2\epsilon^2\sigma^2}{d} + \frac{2\epsilon^2}{d(d-1)}\left[|x_K|^2 - \sigma^2 d + \sigma^2 - W^2\right] + O(\epsilon^3)$$

By the same decomposition and a bit more work,

$$\begin{aligned}
&\mathbb{E}\left[(W_\epsilon - W)^2 \middle| W\right] \\
&\quad = \frac{2\epsilon^2\sigma^2}{d} + \frac{2\epsilon^2}{d(d-1)}\left[|x_K|^2 - \sigma^2 d + \sigma^2 - W^2\right] + O(\epsilon^3) \\
&\quad = 2\lambda(\epsilon)\sigma^2\left(1 + \left[\frac{1}{\sigma^2(d-1)}\left(|x_K|^2 - \sigma^2 d + \sigma^2 - W^2\right)\right]\right) + O(\epsilon^3).
\end{aligned}$$

By the same decomposition and a bit more work,

$$\begin{aligned}
&\mathbb{E}\left[(W_\epsilon - W)^2 \middle| W\right] \\
&\quad = \frac{2\epsilon^2\sigma^2}{d} + \frac{2\epsilon^2}{d(d-1)}\left[|x_K|^2 - \sigma^2 d + \sigma^2 - W^2\right] + O(\epsilon^3) \\
&\quad = 2\lambda(\epsilon)\sigma^2\left(1 + \underbrace{\left[\frac{1}{\sigma^2(d-1)}\left(|x_K|^2 - \sigma^2 d + \sigma^2 - W^2\right)\right]}_{E}\right) + O(\epsilon^3)
\end{aligned}$$

By the same decomposition and a bit more work,

$$\begin{aligned}
&\mathbb{E}\left[(W_\epsilon - W)^2 \middle| W\right] \\
&\quad = \frac{2\epsilon^2\sigma^2}{d} + \frac{2\epsilon^2}{d(d-1)}\left[|x_K|^2 - \sigma^2 d + \sigma^2 - W^2\right] + O(\epsilon^3) \\
&\quad = 2\lambda(\epsilon)\sigma^2\left(1 + \underbrace{\left[\frac{1}{\sigma^2(d-1)}\left(|x_K|^2 - \sigma^2 d + \sigma^2 - W^2\right)\right]}_{E}\right) + O(\epsilon^3)
\end{aligned}$$

Recall: $d_{TV}(W, \sigma Z) \leq \mathbb{E}\left|E\right|$.

By the same decomposition and a bit more work,

$$\begin{aligned}
&\mathbb{E}\left[(W_\epsilon - W)^2 \middle| W\right] \\
&\quad = \frac{2\epsilon^2\sigma^2}{d} + \frac{2\epsilon^2}{d(d-1)}\left[|x_K|^2 - \sigma^2 d + \sigma^2 - W^2\right] + O(\epsilon^3) \\
&\quad = 2\lambda(\epsilon)\sigma^2\left(1 + \underbrace{\left[\frac{1}{\sigma^2(d-1)}\left(|x_K|^2 - \sigma^2 d + \sigma^2 - W^2\right)\right]}_{E}\right) + O(\epsilon^3)
\end{aligned}$$

Recall: $d_{TV}(W, \sigma Z) \leq \mathbb{E}\left|E\right|$.

$$\implies d_{TV}(W, \sigma Z) \leq \frac{1}{d-1}\mathbb{E}\left|\frac{|x_K|^2}{\sigma^2} - d + 1 - \frac{X^2}{\sigma^2}\right|$$

By the same decomposition and a bit more work,

$$\begin{aligned}
&\mathbb{E}\left[(W_\epsilon - W)^2 \middle| W\right] \\
&\quad = \frac{2\epsilon^2\sigma^2}{d} + \frac{2\epsilon^2}{d(d-1)}\left[|x_K|^2 - \sigma^2 d + \sigma^2 - W^2\right] + O(\epsilon^3) \\
&\quad = 2\lambda(\epsilon)\sigma^2\left(1 + \underbrace{\left[\frac{1}{\sigma^2(d-1)}\left(|x_K|^2 - \sigma^2 d + \sigma^2 - W^2\right)\right]}_{E}\right) + O(\epsilon^3)
\end{aligned}$$

Recall: $d_{TV}(W, \sigma Z) \leq \mathbb{E}\left|E\right|$.

$$\begin{aligned}
\implies d_{TV}(W, \sigma Z) &\leq \frac{1}{d-1}\mathbb{E}\left|\frac{|x_K|^2}{\sigma^2} - d + 1 - \frac{X^2}{\sigma^2}\right| \\
&\leq \frac{1}{d-1}\left[\frac{1}{n}\sum_{i-1}^{n}\left|\frac{|x_i|^2}{\sigma^2} - d\right| + 2\right]
\end{aligned}$$

By the same decomposition and a bit more work,

$$\begin{aligned}
&\mathbb{E}\left[(W_\epsilon - W)^2 \middle| W\right] \\
&\quad = \frac{2\epsilon^2\sigma^2}{d} + \frac{2\epsilon^2}{d(d-1)}\left[|x_K|^2 - \sigma^2 d + \sigma^2 - W^2\right] + O(\epsilon^3) \\
&\quad = 2\lambda(\epsilon)\sigma^2\left(1 + \underbrace{\left[\frac{1}{\sigma^2(d-1)}\left(|x_K|^2 - \sigma^2 d + \sigma^2 - W^2\right)\right]}_{E}\right) + O(\epsilon^3)
\end{aligned}$$

Recall: $d_{TV}(W, \sigma Z) \leq \mathbb{E}\left|E\right|$.

$$\begin{aligned}
\implies d_{TV}(W, \sigma Z) &\leq \frac{1}{d-1}\mathbb{E}\left|\frac{|x_K|^2}{\sigma^2} - d + 1 - \frac{X^2}{\sigma^2}\right| \\
&\leq \frac{1}{d-1}\left[\frac{1}{n}\sum_{i-1}^{n}\left|\frac{|x_i|^2}{\sigma^2} - d\right| + 2\right] = \frac{A+2}{d-1}.
\end{aligned}$$

Leveraging to a stronger theorem – the concentration of measure phenomenon

Leveraging to a stronger theorem – the concentration of measure phenomenon

The idea of concentration of measure is that a well-behaved function of a point on the sphere is almost constant, and that constant can be taken to be the average value of the function.

Leveraging to a stronger theorem – the concentration of measure phenomenon

The idea of concentration of measure is that a well-behaved function of a point on the sphere is almost constant, and that constant can be taken to be the average value of the function. More specifically:

Theorem (Lévy's lemma)

There are universal constants C, c such that if $f : \mathbb{S}^{d-1} \to \mathbb{R}$ is Lipschitz with Lipschitz constant L (with respect to either the geodesic distance or the Euclidean distance on $\mathbb{S}^{d-1} \subseteq \mathbb{R}^d$),

$$\mathbb{P}\left[\left|f(x) - \mathbb{E}f\right| > \epsilon\right] \leq C \exp\left(-\frac{c\epsilon^2 d}{L^2}\right).$$

Define a function $F : \mathbb{S}^{d-1} \to \mathbb{R}$ by

$$\begin{aligned} F(\theta) &:= d_{BL}\big(W(\theta), \sigma Z\big) \\ &= \sup_{\|f\|_{BL} \le 1} \left| \frac{1}{n} \sum_{i=1}^{n} f\big(\langle \theta, x_i \rangle \big) - \mathbb{E} f(\sigma Z) \right|. \end{aligned}$$

Define a function $F : \mathbb{S}^{d-1} \to \mathbb{R}$ by

$$F(\theta) := d_{BL}\big(W(\theta), \sigma Z\big) = \sup_{\|f\|_{BL} \le 1} \left| \frac{1}{n} \sum_{i=1}^{n} f\big(\langle \theta, x_i \rangle \big) - \mathbb{E} f(\sigma Z) \right|.$$

It is not hard to show that the Lipschitz constant of F is bounded by $\sqrt{B}$, thus

$$\mathbb{P}\left[\left| d_{BL}\big(W(\theta), \sigma Z\big) - \mathbb{E} d_{BL}\big(W(\theta), \sigma Z\big) \right| > \epsilon \right] \le C e^{-\frac{cd\epsilon^2}{B}},$$

where $\mathbb{E} d_{BL}\big(W(\theta), \sigma Z\big)$ is the mean with respect to θ of the function $d_{BL}\big(W(\theta), \sigma Z\big)$ of θ.

BL-distance as the supremum of a stochastic process

BL-distance as the supremum of a stochastic process

Let $f : \mathbb{R}^k \to \mathbb{R}$ have $\|f\|_{BL} \leq 1$, and define the stochastic process $\{X_f\}$ indexed by such f by

$$\begin{aligned} X_f(\theta) &:= \left| \mathbb{E}_K f\big(W(\theta)\big) - \mathbb{E} f\big(\sigma Z\big) \right| \\ &= \left| \frac{1}{n} \sum_{i=1}^{n} f\big(\langle \theta, x_i \rangle \big) - \mathbb{E} f\big(\sigma Z\big) \right|. \end{aligned}$$

BL-distance as the supremum of a stochastic process

Let $f : \mathbb{R}^k \to \mathbb{R}$ have $\|f\|_{BL} \leq 1$, and define the stochastic process $\{X_f\}$ indexed by such f by

$$X_f(\theta) := \left| \mathbb{E}_K f\big(W(\theta)\big) - \mathbb{E} f\big(\sigma Z\big) \right|$$
$$= \left| \frac{1}{n} \sum_{i=1}^{n} f\big(\langle \theta, x_i \rangle \big) - \mathbb{E} f\big(\sigma Z\big) \right|.$$

Then $d_{BL}\big(W(\theta), \sigma Z\big) = \sup_f X_f$, and we are interested in $\mathbb{E} \sup_f X_f$.

Essentially the same measure-concentration argument as above shows that for $f : \mathbb{R} \to \mathbb{R}$ with $\|f\|_{BL} \leq L$, the function

$$G_f(\theta) := \mathbb{E}_K f\big(W(\theta)\big) = \frac{1}{n} \sum_{i=1}^{n} f\big(\langle \theta, x_i \rangle \big)$$

is concentrated (with respect to Haar measure on $\mathbb{S}^{d-1}$) about its mean:

Essentially the same measure-concentration argument as above shows that for $f : \mathbb{R} \to \mathbb{R}$ with $\|f\|_{BL} \leq L$, the function

$$G_f(\theta) := \mathbb{E}_K f\big(W(\theta)\big) = \frac{1}{n} \sum_{i=1}^{n} f\big(\langle \theta, x_i \rangle\big)$$

is concentrated (with respect to Haar measure on $\mathbb{S}^{d-1}$) about its mean:

$$\mathbb{P}\left[|G_f(\theta) - \mathbb{E}(G_f)| > \epsilon\right] \leq Ce^{-\frac{cd\epsilon^2}{L^2 B}},$$

for some constants C, c.

$$\mathbb{P}\left[|G_f(\theta) - \mathbb{E}(G_f)| > \epsilon\right] \leq Ce^{-\frac{cd\epsilon^2}{L^2 B}}$$

$$\mathbb{P}\left[|G_f(\theta) - \mathbb{E}(G_f)| > \epsilon\right] \leq Ce^{-\frac{cd\epsilon^2}{L^2 B}}$$

Observe that $\mathbb{E}G_f(\theta) = \mathbb{E}f\big(W(\theta)\big)$, and recall that

$$\Big|\mathbb{E}f\big(W(\theta)\big) - \mathbb{E}f\big(\sigma Z\big)\Big| \leq \frac{A+2}{d-1};$$

it follows that there are constants C and c such that for $\epsilon > 0$, $\mathbb{P}\left[X_f > \epsilon\right] \leq Ce^{-\frac{cd\epsilon^2}{BL^2}}$.

$$\mathbb{P}\left[\left|G_f(\theta) - \mathbb{E}(G_f)\right| > \epsilon\right] \leq Ce^{-\frac{cd\epsilon^2}{L^2 B}}$$

Observe that $\mathbb{E}G_f(\theta) = \mathbb{E}f\big(W(\theta)\big)$, and recall that

$$\left|\mathbb{E}f\big(W(\theta)\big) - \mathbb{E}f\big(\sigma Z\big)\right| \leq \frac{A+2}{d-1};$$

it follows that there are constants C and c such that for $\epsilon > 0$, $\mathbb{P}\left[X_f > \epsilon\right] \leq Ce^{-\frac{cd\epsilon^2}{BL^2}}$.

It's not hard to see that $\left|X_f - X_g\right| \leq X_{f-g}$, and so

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

$$\mathbb{P}\left[|G_f(\theta) - \mathbb{E}(G_f)| > \epsilon\right] \leq Ce^{-\frac{cd\epsilon^2}{L^2 B}}$$

Observe that $\mathbb{E}G_f(\theta) = \mathbb{E}f\big(W(\theta)\big)$, and recall that

$$\Big|\mathbb{E}f\big(W(\theta)\big) - \mathbb{E}f\big(\sigma Z\big)\Big| \leq \frac{A+2}{d-1};$$

it follows that there are constants C and c such that for $\epsilon > 0$, $\mathbb{P}\left[X_f > \epsilon\right] \leq Ce^{-\frac{cd\epsilon^2}{BL^2}}$.

It's not hard to see that $\big|X_f - X_g\big| \leq X_{f-g}$, and so

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

that is, $\{X_f\}$ satisfies a sub-Gaussian increment condition with respect to the distance $d(f,g) := \frac{\sqrt{B}\|f-g\|_{BL}}{\sqrt{2cd}}$.

Dudley's entropy bound

Dudley's entropy bound

Theorem (Dudley)

If a stochastic process $\{X_t\}_{t\in T}$ *satisfies the a sub-Gaussian increment condition*

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

then

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

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

The covering number

The covering number

So what we now need is an estimate on $N(T, d, \epsilon)$ for $T = \{f : \mathbb{R} \to \mathbb{R} : \|f\|_{BL} \leq 1\}$ and $d(f, g) = \frac{\sqrt{B}\|f-g\|_{BL}}{\sqrt{2cd}}$.

The covering number

So what we now need is an estimate on $N(T, d, \epsilon)$ for $T = \{f : \mathbb{R} \to \mathbb{R} : \|f\|_{BL} \leq 1\}$ and $d(f, g) = \frac{\sqrt{B}\|f-g\|_{BL}}{\sqrt{2cd}}$.

The problem of course is that $N(T, d, \epsilon) = \infty$ for this choice of T, d. But we can still use Dudley's bound, together with a truncation argument (and a smoothing argument).

The covering number

So what we now need is an estimate on $N(T, d, \epsilon)$ for $T = \{f : \mathbb{R} \to \mathbb{R} : \|f\|_{BL} \leq 1\}$ and $d(f, g) = \frac{\sqrt{B}\|f-g\|_{BL}}{\sqrt{2cd}}$.

The problem of course is that $N(T, d, \epsilon) = \infty$ for this choice of T, d. But we can still use Dudley's bound, together with a truncation argument (and a smoothing argument).

Estimating the covering number of $C_1(\mathcal{X})$ for $\mathcal{X} \subseteq \mathbb{R}$ compact, with respect to d as above, then making truncation and smoothing arguments yields:

$$\mathbb{E}d_{BL}\big(W(\theta), \sigma Z\big) \leq \frac{CB}{d^{\frac{2}{9}}}.$$

Summary

Summary

We now have the following:

Summary

We now have the following:

- $\mathbb{P}\left[\left|d_{BL}\left(W(\theta), \sigma Z\right) - \mathbb{E}d_{BL}\left(W(\theta), \sigma Z\right)\right| > \epsilon\right] \leq \sqrt{\frac{\pi}{2}} e^{-\frac{d\epsilon^2}{8B}}$

Summary

We now have the following:

- $\mathbb{P}\left[\left|d_{BL}\left(W(\theta), \sigma Z\right) - \mathbb{E}d_{BL}\left(W(\theta), \sigma Z\right)\right| > \epsilon\right] \leq \sqrt{\frac{\pi}{2}} e^{-\frac{d\epsilon^2}{8B}}$
- $\mathbb{E}d_{BL}\left(W(\theta), \sigma Z\right) \leq \frac{CB}{d^{\frac{2}{9}}}.$

Summary

We now have the following:

- $\mathbb{P}\left[\left|d_{BL}\left(W(\theta), \sigma Z\right) - \mathbb{E}d_{BL}\left(W(\theta), \sigma Z\right)\right| > \epsilon\right] \leq \sqrt{\frac{\pi}{2}} e^{-\frac{d\epsilon^2}{8B}}$
- $\mathbb{E}d_{BL}\left(W(\theta), \sigma Z\right) \leq \frac{CB}{d^{\frac{2}{9}}}.$

Summary

We now have the following:

- $\mathbb{P}\left[\left|d_{BL}\left(W(\theta), \sigma Z\right) - \mathbb{E}d_{BL}\left(W(\theta), \sigma Z\right)\right| > \epsilon\right] \leq \sqrt{\frac{\pi}{2}} e^{-\frac{d\epsilon^2}{8B}}$
- $\mathbb{E}d_{BL}\left(W(\theta), \sigma Z\right) \leq \frac{CB}{d^{\frac{2}{9}}}$.

In particular, there are constants C, c such that

$$\mathbb{P}\left[d_{BL}\left(W(\theta), \sigma Z\right) > \epsilon\right] \leq Ce^{-\frac{cd\epsilon^2}{B}},$$

and this bound is interesting for $\epsilon > \frac{CB}{d^{\frac{2}{9}}}$.

Higher dimensional projections

Higher dimensional projections

The approach described here can be adapted to treat the case of higher-dimensional projections of the data vectors x_i.

Higher dimensional projections

The approach described here can be adapted to treat the case of higher-dimensional projections of the data vectors x_i. That is, one could ask about projections of the d-dimensional data onto a random k-dimensional subspace for some $k \leq d$;

Higher dimensional projections

The approach described here can be adapted to treat the case of higher-dimensional projections of the data vectors x_i. That is, one could ask about projections of the d-dimensional data onto a random k-dimensional subspace for some $k \leq d$; one could further ask if k can grow with d and, if so, how?

Random subspaces

Random subspaces

The first step is to describe a random k-dimensional subspace of $\mathbb{R}^d$, which can be done in terms of the Stiefel manifold.

Random subspaces

The first step is to describe a random k-dimensional subspace of $\mathbb{R}^d$, which can be done in terms of the Stiefel manifold.
Definition:The Stiefel manifold $\mathfrak{W}_{d,k}$ is defined by

$$\mathfrak{W}_{d,k} := \{(\theta_1, \ldots, \theta_k) : |\theta_i| = 1, \langle \theta_i, \theta_j \rangle = \delta_{ij}\},$$

with rotation-invariant measure described by choosing θ_1 uniformly in $\mathbb{S}^{d-1}$ and choosing θ_i uniformly in the orthogonal complement of $\theta_1, \ldots, \theta_{i-1}$ for $i > 1$.

Random subspaces

The first step is to describe a random k-dimensional subspace of $\mathbb{R}^d$, which can be done in terms of the Stiefel manifold.
Definition:The Stiefel manifold $\mathfrak{W}_{d,k}$ is defined by

$$\mathfrak{W}_{d,k} := \{(\theta_1, \ldots, \theta_k) : |\theta_i| = 1, \langle \theta_i, \theta_j \rangle = \delta_{ij}\},$$

with rotation-invariant measure described by choosing θ_1 uniformly in $\mathbb{S}^{d-1}$ and choosing θ_i uniformly in the orthogonal complement of $\theta_1, \ldots, \theta_{i-1}$ for $i > 1$.

When choosing a random k-dimensional projection of the data, what is meant is to project onto the span of a random point of the Stiefel manifold.

Adaptations of the proof scheme

Adaptations of the proof scheme

1. The analogous averaged theorem, with $W := (\langle \theta_1, X_K \rangle, \ldots, \langle \theta_k, X_K \rangle)$ can be proved as in the one-dimensional case, using a multivariate version of the abstract normal approximation theorem.

Adaptations of the proof scheme

1. The analogous averaged theorem, with $W := (\langle \theta_1, x_K \rangle, \ldots, \langle \theta_k, x_K \rangle)$ can be proved as in the one-dimensional case, using a multivariate version of the abstract normal approximation theorem.
2. The concentration of measure phenomenon occurs on the Stiefel manifold as well (with the same constants), and thus that part of the proof goes through essentially unchanged.

Adaptations of the proof scheme

1. The analogous averaged theorem, with $W := (\langle \theta_1, x_K \rangle, \ldots, \langle \theta_k, x_K \rangle)$ can be proved as in the one-dimensional case, using a multivariate version of the abstract normal approximation theorem.
2. The concentration of measure phenomenon occurs on the Stiefel manifold as well (with the same constants), and thus that part of the proof goes through essentially unchanged.
3. One can also apply Dudley's entropy bound to the analogous stochastic process

$$X_f(\theta) := \left| \frac{1}{n} \sum_{i=1}^{n} f\big(\langle \theta_1, x_i \rangle, \ldots, \langle \theta_k, x_i \rangle \big) - \mathbb{E} f\big(\sigma Z\big) \right|.$$

The covering number used in the bound will of course depend on k.

A multivariate theorem

A multivariate theorem

Theorem (M)

Let $x_1, \dots, x_n$ be deterministic vectors in $\mathbb{R}^d$, and suppose that A and B are defined by $\frac{1}{n}\sum_{i=1}^n \left|\sigma^{-2}|x_i|^2 - d\right| =: A$ and $\sup_{\theta\in\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^n \langle\theta, x_i\rangle^2 =: B$.

A multivariate theorem

Theorem (M)

Let $x_1, \ldots, x_n$ be deterministic vectors in $\mathbb{R}^d$, and suppose that A and B are defined by $\frac{1}{n}\sum_{i=1}^n \left|\sigma^{-2}|x_i|^2 - d\right| =: A$ and $\sup_{\theta\in\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^n \langle\theta, x_i\rangle^2 =: B$. Let θ be a random point of the Stiefel manifold $\mathfrak{W}_{d,k}$, and let X_θ be the random variable on $\mathbb{R}^k$ which is uniformly distributed over the n points $\left\{\left(\langle\theta_1, x_i\rangle, \ldots, \langle\theta_k, x_i\rangle\right)\right\}_{i=1}^n$.

A multivariate theorem

Theorem (M)

Let $x_1, \ldots, x_n$ be deterministic vectors in $\mathbb{R}^d$, and suppose that A and B are defined by $\frac{1}{n}\sum_{i=1}^n \left|\sigma^{-2}|x_i|^2 - d\right| =: A$ and $\sup_{\theta\in\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^n \langle\theta, x_i\rangle^2 =: B$. Let θ be a random point of the Stiefel manifold $\mathfrak{W}_{d,k}$, and let X_θ be the random variable on $\mathbb{R}^k$ which is uniformly distributed over the n points $\left\{\left(\langle\theta_1, x_i\rangle, \ldots, \langle\theta_k, x_i\rangle\right)\right\}_{i=1}^n$. Then there are constants C, c such that

$$\mathbb{P}\left[d_{BL}\left(X_\theta, \sigma Z\right) > \epsilon\right] \leq Ce^{-\frac{cd\epsilon^2}{B}},$$

and this bound is non-trivial for $\epsilon > \frac{C^k B}{d^{\frac{2}{5k+4}}}$.

Thank you.