IV - High-Dimensional Geometry and Some Applications

Math 728 D - Machine Learning & Data Science - Spring 2019

Contents

1. Volumes in High Dimensions
2. Concentration of Measure
3. Dimension Reduction
4. First Applications to Simple Estimation Problems
   - Separating Gaussians
   - Fitting a Gaussian
   - Gaussian Mixtures and Clustering

Effect of Shrinking

Consider $A \subset \mathbb{R}^d$ measurable, $\epsilon \in (0,1)$, $(1-\epsilon)A := \{(1-\epsilon)\mathbf{x} : \mathbf{x} \in A\}$; let

$$\mathrm{vol}(A) = \mathrm{vol}_d(A) := \int_A \chi_A(\mathbf{x})d\mathbf{x} \quad \text{(volume of } A\text{)}.$$

Then

$$\mathrm{vol}\big((1-\epsilon)A\big) = (1-\epsilon)^d \mathrm{vol}(A). \tag{2.1}$$

Argument: this holds for any d-dimensional cube (induction on d); cover A by cubes of smaller and smaller size; additivity of the volumes of the cubes + each cube shrinks by factor $(1-\epsilon)^d$, measurability of A (see Lecture II, page 6) $\rightsquigarrow$ (2.1).

Hence

$$\frac{\mathrm{vol}\big((1-\epsilon)A\big)}{\mathrm{vol}(A)} = (1-\epsilon)^d \le e^{-\epsilon d}, \tag{2.2}$$

i.e., such fractions decay exponentially when d increases.

The Euclidean Ball/Sphere

Define

$$B_d := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \le 1\} \quad S_d := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\} = \partial B_d.$$

We are interested in the quantities

$$V(d) := \mathrm{vol}_d(B_d), \quad A(d) := \mathrm{vol}_{d-1}(S_d).$$

Cartesian Coordinates:

$$V(d) = \int_{x_1=-1}^{x_1=1} \int_{x_2=-\sqrt{1-x_1^2}}^{x_2=\sqrt{1-x_1^2}} \cdots \int_{x_d=-\sqrt{1-x_1^2-\cdots-x_{d-1}^2}}^{x_d=\sqrt{1-x_1^2-\cdots-x_{d-1}^2}} dx_d\, dx_{d-1} \cdots dx_2\, dx_1,$$

or, in radial coordinates:

$$V(d) = \int_{S_d} \int_{r=0}^{1} r^{d-1}\, dr\, dA = \int_{S_d} dA \int_{r=0}^{1} r^{d-1}\, dr = \frac{A(d)}{d}.$$

How to compute $A(d)$?

The Euclidean Ball/Sphere

Compute instead

$$G(d) := \int_{\mathbb{R}^d} e^{-\|\mathbf{x}\|_2^2} d\mathbf{x} = \prod_{j=1}^{d} \int_{\mathbb{R}} e^{-x_j^2} dx_j = \pi^{\frac{d}{2}} \quad (\text{since } \int_{\mathbb{R}} e^{-x^2} dx = \sqrt{\pi}). \tag{2.3}$$

Calculate $G(d)$ using polar coordinates ($e^{-\|\mathbf{x}\|_2^2} = e^{-r^3}$ for $\mathbf{x}$ in the sphere with radius r)

$$G(d) = \int_{S_d} dA \int_0^{\infty} e^{-r^2} r^{d-1} dr = A(d) \int_0^{\infty} e^{-r^2} r^{d-1} dr = A(d) \frac{1}{2} \Gamma\left(\frac{d}{2}\right). \tag{2.4}$$

where $\Gamma(x) := \int_0^{\infty} e^{-z} z^{x-1} dx$ is the Gamma-function (generalizing the factorial $\Gamma(n+1) = n!$).

(2.3), (2.4) $\Rightarrow$

$$A(d) = 2\pi^{\frac{d}{2}} \Gamma\left(\frac{d}{2}\right)^{-1} \quad \rightsquigarrow \tag{2.5}$$

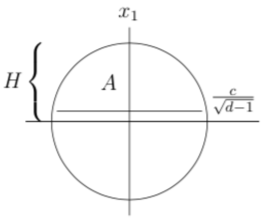
Remark 1

$$V(d) = \frac{2}{d} \pi^{\frac{d}{2}} \Gamma\left(\frac{d}{2}\right)^{-1}, \quad A(d) = 2\pi^{\frac{d}{2}} \Gamma\left(\frac{d}{2}\right)^{-1}.$$

Compare with the volume 2^d von the ℓ_∞^d ball $[-1,1]^d$; what is the probabilityof uniform samples over $[-1,1]^d$ to land in B_d?

Concentration of Measure

Most of the measure of B_d is concentrated for large d in a slab around an equator. W.l.o.g. let $\mathbf{e}^1$ be the north pole.



Theorem 2

Let $c \geq 1$ and

$$SI(c) = \{\mathbf{x} \in B_d : |x_1| \leq c/\sqrt{d-1}\}.$$

Then, for $d \geq 3$

$$\frac{\mathrm{vol}(SI(c))}{\mathrm{vol}(B_d)} \geq 1 - \frac{2}{c} e^{-c^2/2}. \tag{3.1}$$

Proof of Theorem 2: Use notation in the above figure. By symmetry, it suffices to show that

$$\frac{\mathrm{vol}(A)}{\mathrm{vol}(H)} \le \frac{2}{c} e^{-c^2/2}. \tag{3.2}$$

Upper bound for $\mathrm{vol}(A)$: Consider a disk at height $x_1 \ge 0$ of (infinitesimally small) width δx_1 whose top face is a $(d-1)$ dimensional ball of radius $\sqrt{1-x_1^2}$. Since the surface area is $V(d-1)(1-x_1^2)^{\frac{d-1}{2}}$ its volume is $\delta x_1 V(d-1)(1-x_1^2)^{\frac{d-1}{2}}$. The volume of A is obtained by adding the volumes of these disks and letting $\delta x_1 \to 0$; $\rightsquigarrow$

$$\mathrm{vol}(A) = \int_{\frac{c}{\sqrt{d-1}}}^{1} V(d-1)(1-x_1^2)^{\frac{d-1}{2}} dx_1 \overset{(1-x)\le e^{-x}}{\le} \int_{\frac{c}{\sqrt{d-1}}}^{\infty} V(d-1) e^{-x_1^2 \frac{d-1}{2}} dx_1$$

$$\overset{\frac{x_1\sqrt{d-1}}{c}\ge 1}{\le} V(d-1)\frac{\sqrt{d-1}}{c} \int_{\frac{c}{\sqrt{d-1}}}^{\infty} x_1 e^{-x_1^2 \frac{d-1}{2}} dx_1.$$

Since $\int_{\frac{c}{\sqrt{d-1}}}^{\infty} x_1 e^{-x_1^2 \frac{d-1}{2}} dx_1 = -(d-1)^{-1} e^{-x_1^2 \frac{d-1}{2}} \Big|_{\frac{c}{\sqrt{d-1}}}^{\infty} = (d-1)^{-1} e^{-c^2/2} \rightsquigarrow$

$$\mathrm{vol}(A) \le \frac{V(d-1)}{c\sqrt{d-1}} e^{-c^2/2}. \tag{3.3}$$

Proof of Theorem 2 continued: Lower bound for $\mathrm{vol}(H)$:

Consider the cylinder ($x_1 = (d-1)^{-1/2}$)

$$C := (0,(d-1)^{-1/2})\times(1-(d-1)^{-1})^{1/2}V(d-1) \quad \rightsquigarrow \quad \mathrm{vol}(C) = \frac{(1-(d-1)^{-1})^{\frac{d-1}{2}}}{\sqrt{d-1}}V(d-1)$$

For $a \geq 1$ one has $(1-x)^a \geq 1-ax$ (note that for $d \geq 3$ one has $a := (d-1)/2 \geq 1$) $\rightsquigarrow$

$$\mathrm{vol}(H) \geq \mathrm{vol}(SI(1)) \geq \mathrm{vol}(C) = \frac{(1-(d-1)^{-1})^{\frac{d-1}{2}}}{\sqrt{d-1}}V(d-1) \geq \frac{\frac{1}{2}}{\sqrt{d-1}}V(d-1).$$

By (3.3)

$$\frac{\mathrm{vol}(A)}{\mathrm{vol}(H)} \leq \frac{\frac{V(d-1)}{c\sqrt{d-1}}e^{-c^2/2}}{\frac{\frac{1}{2}}{\sqrt{d-1}}V(d-1)} = \frac{2}{c}e^{-c^2/2}.$$

□

Near Orthogonality

Consequences:

Theorem 3

Draw n points $\mathbf{x}^1, \ldots, \mathbf{x}^n$ *at random (uniform distribution) from the unit ball* B_d*: then with probability at least* $1 - 1/n$*, one has*

1. $\|\mathbf{x}^i\|_2 \geq 1 - \frac{2 \log n}{d}$ *for all* $i \in \{1, 2, \ldots, n\}$ *and*
2. $|\mathbf{x}^i \cdot \mathbf{x}^j| \leq \frac{\sqrt{6 \log n}}{\sqrt{d-1}}$ *for all* $i \neq j$*.*

Comments:

- (1) says that n randomly drawn points accumulate with the higher probability near the boundary S_d of B_d the larger d.
- (2) says that the inner product of any two of the n randomly drawn points is close to zero with high probability when d gets large. In view of (1) this actually means that the larger d "the more orthogonal" get pairs of randomly drawn points (recall: $\frac{|\mathbf{x} \cdot \mathbf{y}|}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} = \cos(\angle(\mathbf{x}, \mathbf{y}))$)
- Theorem 3 quantifies the earlier observations derived from the Law of Large Numbers in Lecture II.
- Estimating probabilities in conjunction with "for all" statements is usually done with the aid of so called union bounds, see next page.

Union Bounds a frequent argument

The Union Bound is a frequently used "argument macro" which is a Boolean inequality and often comes in the following form.

Remark 4

Let $X_j \sim (\mathcal{X}, \mathcal{B}, P)$, $j \in \mathcal{I}$. Assume that for some $A \in \mathcal{B}$ and each X_j one knows that $\mathrm{Prob}\Big(X_j \notin A\Big) \le \delta_j$, $j \in \mathcal{I}$. Then

$$\mathrm{Prob}\Big(\forall j \in \mathcal{I} : X_j \in A\big) \ge 1 - \sum_{j \in \mathcal{I}} \delta_j. \tag{3.4}$$

In detail:

$$\mathrm{Prob}\big(\forall j \in \mathcal{I} : X_j \in A\big) = 1 - \mathrm{Prob}\Big(\exists j \text{ such that } X_j \notin A\Big). \tag{3.5}$$

Defining the event $A_j = \{\omega \in \Omega : X_j \notin A\}$,

$$\begin{aligned} \mathrm{Prob}\Big(\exists j \in \mathcal{I} \text{ such that } X_j \notin A\Big) &= \mathrm{Prob}\Big(\mathrm{or}_{j\in\mathcal{I}}(X_j \notin A)\Big) = P\Big(\bigcup_{j\in\mathcal{I}} A_j\Big) \le \sum_{j\in\mathcal{I}} P(A_j) \\ &= \sum_{j\in\mathcal{I}} \mathrm{Prob}\big(X_j \notin A\big) \le \sum_{j\in\mathcal{I}} \delta_j. \end{aligned} \tag{3.6}$$

(3.6) + (3.5) $\Rightarrow$ (3.4). □

Proof of Theorem 3: ad (1): Let **X** be uniformly distributed over B_d. By (2.2)

$$\mathrm{Prob}\Big(\|\mathbf{X}\|_2 < 1-\epsilon\Big) \le \frac{\mathrm{vol}((1-\epsilon)B_d)}{\mathrm{vol}(B_d)} \le e^{-\epsilon d}.$$

Thus, for each fixed $i \in \{1,\ldots,n\}$

$$\mathrm{Prob}\Big(\|\mathbf{X}^i\|_2 < 1-\frac{2\log n}{d}\Big) \le e^{-\left(\frac{2\log n}{d}\right)d} = \frac{1}{n^2}.$$

Hence

$$\begin{aligned}
&\mathrm{Prob}\Big(\exists\, i \text{ s.t. } \|\mathbf{X}^i\|_2 < 1-\frac{2\log n}{d}\Big)\\
&\le P\Big(\Big\{\mathbf{X}^1 : \|\mathbf{X}^1\|_2 < 1-\frac{2\log n}{d}\Big\} \cup \cdots \cup \Big\{\mathbf{X}^n : \|\mathbf{X}^n\|_2 < 1-\frac{2\log n}{d}\Big\}\Big)\\
&\le \frac{n}{n^2} = \frac{1}{n} \quad \Rightarrow \quad \mathrm{Prob}\Big(\forall\, i\, \|\mathbf{X}^i\|_2 \ge 1-\frac{2\log n}{d}\Big) \ge 1-\frac{1}{n} \ \rightsquigarrow\ (1),
\end{aligned}$$

where we have used the union bound, see Remark 4 with $A_j \leftrightarrow \Big(\|\mathbf{X}^j\|_2 \ge 1-\frac{2\log n}{d}\Big)$.

Proof of Theorem 3 continued: ad (2): For any fixed among the $\binom{n}{2}$ pairs (i,j) we let $\mathbf{X}^i = X_1\mathbf{e}^1$ have the direction of the north pole, i.e., $\|\mathbf{X}^i\|_2 = |X_1^i|$. By Theorem 2,

$$\mathrm{Prob}\Big(|X_1^j| > \frac{c}{\sqrt{d-1}}\Big) = \frac{\mathrm{vol}(B_d \setminus Sl(c))}{\mathrm{vol}(B_d)} \le \frac{2}{c}e^{-c^2/2}.$$

Therefore, taking $c = \sqrt{6\log n}$, the probability that the projection of $\mathbf{X}^j$ to the north pole-direction is more than $\sqrt{\frac{6\log n}{d-1}}$ can be bounded by (since $6\log 2 > 4$)

$$\mathrm{Prob}\Big(|X_1^j| > \sqrt{\frac{6\log n}{d-1}}\Big) \le \frac{2}{\sqrt{6\log n}}e^{-\frac{6\log n}{2}} \le n^{-3}.$$

The same union bound (Remark 4) implies that the probability, that for some pair (i,j) one has $|\mathbf{X}^i\cdot\mathbf{X}^j| > \sqrt{\frac{6\log n}{d-1}}$, is bounded by $\binom{n}{2}\cdot n^{-3} \le \frac{1}{2n}$. $\Rightarrow$ (2) □

Uniform Random Sampling from the Sphere S_d

Let $X_j \sim \mathcal{N}(0,1)$, $j = 1,\dots,d$, independent standard Gaussians; $\rightsquigarrow$ joint density

$$p_d(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0},\mathbf{I}) = \prod_{j=1}^{d} \mathcal{N}(x_j|0,1) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{x_1^2+\cdots+x_d^2}{2}} = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\|\mathbf{x}\|_2^2}.$$

It is easy to sample according to $\mathcal{N}(x_j|0,1)$ - why? $\rightsquigarrow$ sample according to p_d $\rightsquigarrow$ $\mathbf{X}$ $\rightsquigarrow$ $\mathbf{Y} = \mathbf{X}/\|\mathbf{X}\|_2$

Note: components of **Y** are no longer independent!

Question: how to sample uniformly from B_d?

Gaussian Annulus Theorem

The next theorem describes where the mass of a spherical Gaussian density in high dimensions is concentrated.

Theorem 5

Let $\mathcal{N}(\mathbf{x}|\mathbf{0},\mathbf{I}) = \prod_{j=1}^{d} \mathcal{N}(x_j|0,1)$ *be the* d*-dimensional standard spherical Gaussian density and* $\mathbf{X} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$*. Then, for any* $\beta \leq \sqrt{d}$

$$\mathrm{Prob}\left(\sqrt{d} - \beta \leq \|\mathbf{X}\|_2 \leq \sqrt{d} + \beta\right) = \int\limits_{\sqrt{d}-\beta\leq\|\mathbf{X}\|_2\leq\sqrt{d}+\beta} \mathcal{N}(\mathbf{x}|\mathbf{0},\mathbf{I})d\mathbf{x} \geq 1 - 3e^{-c\beta^2}, \tag{3.7}$$

where c *is a fixed positive constant.*

Intuition: $\mathbf{X} \sim \mathcal{N}(\mathbf{x}|\mathbf{0},\mathbf{I}) \rightsquigarrow \mathbb{E}[\|\mathbf{X}\|_2^2] = \sum_{j=1}^{d} \mathbb{E}[X_j^2] = \sum_{j=1}^{d} \mathrm{var}[X_j] = d$. Thus the expected distance of a point, drawn from $\mathcal{N}(\mathbf{0},\mathbf{I})$, from the origin (the mean) is $\sqrt{d}$. Theorem 5 says that randomly drawn points indeed concentrat tightly around the sphere of radius $\sqrt{d}$.

Proof of Theorem 5: Note

$$\sqrt{d} - \beta \le \|\mathbf{X}\|_2 \le \sqrt{d} + \beta \quad \Leftrightarrow \quad |\|\mathbf{X}\|_2 - \sqrt{d}| \le \beta \tag{3.8}$$

$\rightsquigarrow$ suffices to prove that $\mathrm{Prob}\Big(|\|\mathbf{X}\|_2 - \sqrt{d}| \ge \beta\Big) \le 3e^{-c\beta^2}$. Multiplication by $\|\mathbf{X}\|_2 + \sqrt{d} \rightsquigarrow$

$$|\|\mathbf{X}\|_2^2 - d| \ge (\|\mathbf{X}\|_2 + \sqrt{d})\beta \ge \beta\sqrt{d} \quad \rightsquigarrow$$

$$\mathrm{Prob}\Big(|\|\mathbf{X}\|_2 - \sqrt{d}| \ge \beta\Big) \le \mathrm{Prob}\Big(|\|\mathbf{X}\|_2^2 - d| \ge \beta\sqrt{d}\Big).$$

Rewrite

$$\|\mathbf{X}\|_2^2 - d = \sum_{j=1}^{d} X_j^2 - d = \sum_{j=1}^{d}(X_j^2 - 1) =: \sum_{j=1}^{d} Y_j \quad \rightsquigarrow \quad \mathbb{E}[Y_j] = \mathbb{E}[X_j^2] - 1 = \mathrm{var}[X_j] - 1 = 0.$$

Goal: estimate

$$\mathrm{Prob}\Big(|\|\mathbf{X}\|_2^2 - d| \ge \beta\sqrt{d}\Big) = \mathrm{Prob}\Big(\Big|\sum_{j=1}^{d} Y_j\Big| \ge \beta\sqrt{d}\Big).$$

To apply Theorem 5 we need to bound the rth moments of Y_j.

Proof of Theorem 5 continued: Bounding $\mathbb{E}[Y_j^r]$ ($Y_j = X_j^2 - 1$): to that end, note

$$|Y_j|^r \leq \begin{cases} 1, & \text{for} \quad |X_j| \leq 1, \\ |X_j|^{2r}, & \text{for} \quad |X_j| \geq 1. \end{cases} \qquad \Rightarrow$$

$$|\mathbb{E}[Y_j^r]| = \mathbb{E}[|Y_j|^r] \leq \mathbb{E}[1 + X_j^{2r}] = 1 + \mathbb{E}[X_j^{2r}] = 1 + \sqrt{\frac{2}{\pi}} \int_0^\infty x^{2r} e^{-x^2/2} dx.$$

To estimate $\sqrt{\frac{2}{\pi}} \int_0^\infty x^{2r} e^{-x^2/2} dx$ use that $\Gamma(y) = \int_0^\infty x^{y-1} e^{-x} dx$:

Change of variables $z := x^2/2 \rightsquigarrow$

$$1 + \sqrt{\frac{2}{\pi}} \int_0^\infty x^{2r} e^{-x^2/2} dx = 1 + \sqrt{\frac{1}{\pi}} \int_0^\infty 2^r z^{r-1/2} e^{-z} dz = 1 + \sqrt{\frac{1}{\pi}} 2^r \Gamma(r - 1/2) \leq 2^r r!.$$

Recall: in Lecture III, Theorem 6 we need the rth moment to be bounded by $\sigma^2 r!$.

$$\mathbb{E}[Y_j] = 0, \rightsquigarrow \operatorname{var}[Y_j] = \mathbb{E}[Y_j^2] \overset{r=2}{\leq} 2^2 \cdot 2 = 8 = \sigma_Y^2.$$

Proof of Theorem 5 continued: So far we have $|\mathbb{E}[Y_j^r]| \leq 2^r r!$ but $2^r r! \not\leq 8^2 r!$ $\rightsquigarrow$ another change of variables: $W_j := Y_j/2$ (Lecture II, (8.6)) $\rightsquigarrow$

$$\mathrm{var}[W_j] = \frac{1}{4}\mathrm{var}[Y_j] \leq 2 = \sigma_W^2, \quad \mathbb{E}[W_j^r] = 2^{-r}\mathbb{E}[Y_j^r] \leq r!.$$

Since

$$\mathrm{Prob}\Big(|\|\mathbf{X}\|_2^2 - d| \geq \beta\sqrt{d}\Big) = \mathrm{Prob}\Big(\Big|\sum_{j=1}^d Y_j\Big| \geq \beta\sqrt{d}\Big) = \mathrm{Prob}\Big(\Big|\sum_{j=1}^d W_j\Big| \geq \frac{\beta\sqrt{d}}{2}\Big),$$

Lecture III, Theorem 6 yields ($a = \frac{\beta\sqrt{d}}{2}$),

$$\mathrm{Prob}\Big(|\|\mathbf{X}\|_2^2 - d| \geq \beta\sqrt{d}\Big) \leq 3e^{-\frac{a^2}{12d2}} = 3e^{-\frac{\beta^2}{12\cdot 8}} = 3e^{-\frac{\beta^2}{96}}.$$

$\rightsquigarrow c = 1/96.$ □

Motivation

- One of the most frequent tasks involving high-dimensional data is nearest-neighbor-search.
- Scenario: given is a database of N points $\mathcal{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^N\} \subset \mathbb{R}^d$, $j = 1, \dots, N$, N, d large; $\mathcal{X}$ is efficiently stored.
- Task: for any query point $\mathbf{x} \in \mathbb{R}^d$ find the nearest (or approximately nearest) neighbor from $\mathcal{X}$.
- Wishlist: the number of queries is typically large $\rightsquigarrow$ the response time (returning the neighbor) should be small; typically a moderately growing function of $\log N$ and $\log d$. Preprocessing time is allowed to be larger, e.g. polynomial in N and d.
- An important preprocessing ingredient is dimension reduction, i.e., the projection of $\mathcal{X} \subset \mathbb{R}^d$ to $\mathbb{R}^k$ with $k \ll d$, while approximately preserving mutual distances.

The next result shows how much the dimension can be reduced and how to find a good projection. It is an application of Theorem 5.

The Johnson-Lindenstrauss-Lemma Random Projections

For $k \leq d$ consider the random matrix

$$\boldsymbol{A} = (A_{i,j})_{i,j=1}^{k,d} \in \mathbb{R}^{k \times d} \quad \text{where} \quad A_{i,j} \sim \mathcal{N}(0,1),\ i,j = 1,\dots,k,d, \text{ drawn independently.} \tag{4.1}$$

Let us denote by $\boldsymbol{A}_i = (a_{i,1},\dots,a_{i,d})$, $i = 1,\dots,k$, the rows of $\boldsymbol{A}$. Note: $\boldsymbol{A}_i \sim \mathcal{N}(\mathbf{0},\mathbf{I})$.

We will see: the mapping $\mathbf{x} \in \mathbb{R}^d \mapsto \boldsymbol{A}\mathbf{x} \in \mathbb{R}^k$ is with high probability (regarding the choice of $\boldsymbol{A}$) near-distance preserving..

Theorem 6

Let $\mathbf{x} \in \mathbb{R}^d$ be fixed and let the random matrix $\boldsymbol{A}$ be given by (4.1). Then

$$\mathrm{Prob}\Big(\big|\|\boldsymbol{A}\mathbf{x}\|_2 - \sqrt{k}\|\mathbf{x}\|_2\big| \geq \epsilon\sqrt{k}\|\mathbf{x}\|_2\Big) \leq 3e^{-ck\epsilon^2}, \tag{4.2}$$

where c is the constant from Theorem 5 and the probability is taken with respect to $\mathcal{N}(\cdot|\mathbf{0},\mathbf{I})^k$.

Remark: Since $\boldsymbol{A}$ is linear, for any fixed $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$ one has

$$\left|\frac{\|k^{-1/2}\boldsymbol{A}(\mathbf{x}-\mathbf{y})\|_2}{\|\mathbf{x}-\mathbf{y}\|_2} - 1\right| \leq \epsilon$$

with probability at least $1 - 3e^{-ck\epsilon^2}$.

Proof of Theorem 6: $\boldsymbol{A}\mathbf{x}$ is the vector with components $\boldsymbol{A}_i \cdot \mathbf{x}$, $i = 1, \dots, k$. Dividing both sides in $\mathrm{Prob}\Big(\big| \|\boldsymbol{A}\mathbf{x}\|_2 - \sqrt{k}\|\mathbf{x}\|_2 \big| \geq \epsilon\sqrt{k}\|\mathbf{x}\|_2\Big)$ by $\|\mathbf{x}\|_2$, we can assume without loss of generality that $\|\mathbf{x}\|_2 = 1$ (the statement is about relative accuracy). By Lecture II, Corollary 18 and (10.9), the sum of independent Gaussians is Gaussian whose variance is the sum of variances. $\rightsquigarrow$

$$\mathrm{var}[\boldsymbol{A}_i \cdot \mathbf{x}] = \sum_{j=1}^{d} x_j^2 \mathrm{var}[A_{i,j}] = \sum_{j=1}^{d} x_j^2 = \|\mathbf{x}\|_2^2 = 1.$$

Hence $\boldsymbol{A}_1 \cdot \mathbf{x}, \dots, \boldsymbol{A}_k \cdot \mathbf{x}$ are independent Gaussian variables $\sim \mathcal{N}(0, 1)$. Hence $\boldsymbol{A}\mathbf{x}$ is a k-dimensional spherical Gaussian random variable with unit variance in each coordinate.

Theorem 5 (with d replaced by k and using (3.8)) $\Rightarrow \mathrm{Prob}\Big(\big| \|\boldsymbol{A}\mathbf{x}\|_2 - \sqrt{k} \big| \geq \epsilon\sqrt{k}\Big) \leq 3e^{-ck\epsilon^3}$. $\square$

The Johnson-Lindenstrauss-Lemma

The JL-Lemma is based on the random projection (4.1): define

$$\mathbf{F}(\mathbf{x}) := \frac{1}{\sqrt{k}} \boldsymbol{A}\mathbf{x}. \tag{4.3}$$

Theorem 7

Given: any $\epsilon \in (0,1)$, $N \in \mathbb{N}$; let $k \geq \frac{3 \log N}{c\epsilon^2}$, where c is the constant from Theorem 5.
Claim: for any set $\mathcal{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^N\} \subset \mathbb{R}^d$, the mapping $\mathbf{F}$, defined by (4.3), satisfies for all pairs $\mathbf{x}^i, \mathbf{x}^j \in \mathcal{X}$

$$(1-\epsilon)\|\mathbf{x}^i - \mathbf{x}^j\|_2 \leq \|\mathbf{F}(\mathbf{x}^i) - \mathbf{F}(\mathbf{x}^j)\|_2 \leq (1+\epsilon)\|\mathbf{x}^i - \mathbf{x}^j\|_2 \tag{4.4}$$

holds with probability at least $1 - \frac{3}{2N}$.

Remarks:

- The reduced dimension k does not depend on the ambient dimension d, but only on the number N of projected points.
- The dependence of k on N is only logarithmic.
- There is a close connection between random projections and the Compressive Sensing paradigm discussed later in the course (if time permits).

Proof of Theorem 7: Fix any pair $\mathbf{x}^i, \mathbf{x}^j \in \mathcal{X}$. By the Random Projection Theorem 6, the probability of $\|\mathbf{F}(\mathbf{x}^i) - \mathbf{F}(\mathbf{x}^j)\|_2 = \|\mathbf{F}(\mathbf{x}^i - \mathbf{x}^j)\|_2$ being outside the interval $\big[(1-\epsilon)\|\mathbf{x}^i - \mathbf{x}^j\|_2, (1+\epsilon)\|\mathbf{x}^i - \mathbf{x}^j\|_2\big]$, is at most $3e^{-ck\epsilon^2}$.

For $k \geq \frac{3 \log N}{c\epsilon^2}$, this probability is at most $3/N^3$. Since there are $\binom{N}{2} < N^2/2$ such pairs, the assertion follows from a union bound, see Remark 4. □

Mixtures of Gaussians - An Example

Gaussian mixtures: are often used to model heterogeneous data coming from multiple sources

Example: The heights of individuals in a fixed age range in a city are being recorded. On average men are taller than women $\rightsquigarrow$ Model:

$$\begin{array}{lll} \text{f-height}: & \mu_1 + X_1, & X_1 \sim \mathcal{N}(0, \sigma_1^2); \\ \text{m-height}: & \mu_2 + X_2, & X_2 \sim \mathcal{N}(0, \sigma_2^2). \end{array} \rightsquigarrow \quad p(x) = w_1 \mathcal{N}(x|\mu_1, \sigma_1^2) + w_2 \mathcal{N}(x|\mu_2, \sigma_2^2), \quad (5.1)$$

where the mixture weights w_1, w_2 represent the proportions of females, males in the city.

Problem: Given access to sample from the density $p(x)$, i.e., heights of individuals without knowing the gender, reconstruct the parameters μ_i, σ_i^2, $i = 1, 2$ for the mixture model (5.8).

Notice: since there are shorter men than some women, given a height, it is not clear whether it comes from a female or male.

One could ask analogous questions for more attributes $X_1, ...X_d$.

In this section: Separate two spherical Gaussians with unit-variance for large d but with well-separated means; later: the case of nearby means.

Separation of Gaussians

Model: $p(\mathbf{x}) = w_1\mathcal{N}(\mathbf{x}|\mu_1, 1) + w_2\mathcal{N}(\mathbf{x}|\mu_2, 1)$, $\mathbf{x} \in \mathbb{R}^d$ (d large), find μ_i, w_i, $i = 1, 2$.

Observation 1: For two independent draws $\mathbf{x}, \mathbf{y}$ from the same $\mathcal{N}(\mathbf{0}, \mathbf{I})$, say, one has

$$\|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{2d} \pm O(1). \tag{5.2}$$

Argument: By Theorem 5, $\mathbf{x}, \mathbf{y}$ are with high probability within an annulus of width $O(1)$ around the sphere with radius $\sqrt{d}$. W.l.o.g. we can rotate the coordinate system to obtain $\mathbf{x} = (\sqrt{d} + O(1))\mathbf{e}^1$. By Theorem 2, with high probability, $|\mathbf{y} \cdot \mathbf{e}^1| \leq \sqrt{d} \cdot O((d-1)^{-1/2}) = O(1)$, i.e., $|\mathbf{x} \cdot \mathbf{y}| = O(\sqrt{d}) \rightsquigarrow$

$$\|\mathbf{x} - \mathbf{y}\|_2^2 = (\mathbf{x} - \mathbf{y})^\top(\mathbf{x} - \mathbf{y}) = \|\mathbf{x}\|_2^2 - 2\mathbf{x} \cdot \mathbf{y} + \|\mathbf{y}\|_2^2 = 2d \pm O(\sqrt{d}) \quad \Rightarrow \quad (5.2).$$

Observation 2: Consider two independent draws $\mathbf{x}$, $\mathbf{y}$ from $\mathcal{N}(\mu_1, \mathbf{I}), \mathcal{N}(\mu_2, \mathbf{I})$, respectively, and set $\Delta := \|\mu_1 - \mu_2\|_2$. Then, with high probability one has

$$\|\mathbf{y} - \mathbf{x}\|_2^2 = \Delta^2 + 2d \pm O(\sqrt{d}). \tag{5.3}$$

Argument: Adding, subtracting μ_1, μ_2 and expanding, yields

$$\|\mathbf{x}-\mathbf{y}\|_2^2 = \|\mathbf{x}-\mu_1\|_2^2 + \|\mathbf{y}-\mu_2\|_2^2 + \Delta^2 + 2(\mathbf{x}-\mu_1)^\top(\mathbf{y}-\mu_2) + 2(\mathbf{x}-\mu_1)^\top(\mu_1-\mu_2) - 2(\mathbf{y}-\mu_2)^\top(\mu_1-\mu_2).$$

By the above argument, the 4th summand is $\pm O(\sqrt{d})$. Consider the slabs S_1, S_2 of width $O(1)$ around the centers μ_1, μ_2, which are perpendicular to $\mu_1 - \mu_2$. As argued above, with high probability $\mathbf{x} \in S_1$, $\mathbf{y} \in S_2$ so that $\mu_1 - \mu_2$ has inner products with $\mathbf{x} - \mu_1$, $\mathbf{y} - \mu_2$ of at most the order $O(\sqrt{d}) \Rightarrow (5.3)$.

Outline of a Simple Separation Algorithm

Rationale: Distance D_1 between two points from the same Gaussian should be smaller than the distance D_2 bewteen two points from different Gaussians, i.e.,

$$D_1 \leq \sqrt{2d} + O(1) \overset{!}{\leq} \sqrt{\Delta^2 + 2d} - O(1) \leq D_2 \quad \Leftrightarrow \quad 2d + O(\sqrt{d}) \leq 2d + \Delta^2.$$

This holds when $\Delta \geq Cd^{1/4}$.

Algorithm:

- Calculate all pairwise distances between the samples;
- Identify the two clusters $\mathcal{C}_s, \mathcal{C}_l$ of small and large pairwise distances; pick a pair $(\mathbf{x}^{i_1}, \mathbf{x}^{i_2})$ from $\mathcal{C}_s$ and fix $\mathbf{x}^{i_1}$; define $\mathcal{C}_{s,1}$ as the set of all points $\mathbf{x}^j$ such that $(\mathbf{x}^{i_1}, \mathbf{x}^j) \in \mathcal{C}_l$ (long distance); these points come from a single Gaussian with high probability;
- the remaining points come from the other one.

One still needs to fit the clustered points to a Gaussian.

Maximum Likelihood Estimator (MLE)

Suppose that $\mathbf{x}^1, \ldots, \mathbf{x}^N$ are i.i.d samples from $\mathbf{X} \sim \mathcal{N}(\mu, \sigma^2 \mathbf{I})$ (spherical Gaussian with center $\mu \in \mathbb{R}^d$)

Goal: estimate μ and σ^2 from these points.

The joint density of the underlying random variables $\mathbf{X}^j, j = 1, \ldots, \mathbf{X}^N$ is the dN-dimensional spherical Gaussian

$$p(\mathbf{x}^1, \ldots, \mathbf{x}^N) := \mathcal{N}(\mathbf{x}^1, \ldots, \mathbf{x}^N | (\mu, \ldots, \mu), \sigma^2 \mathbf{I}_{dN}) = \frac{1}{(2\pi\sigma^2)^{\frac{dN}{2}}} e^{-\frac{1}{2\sigma^2}\left(\|\mathbf{x}^1 - \mu\|_2^2 + \cdots + \|\mathbf{x}^N - \mu\|_2^2\right)}.$$

The Maximum Likelihood Estimator (MLE) determines estimates μ_{ML}, σ^2_{ML} by maximizing this joint density for the given data $\mathbf{x}^1, \ldots, \mathbf{x}^N$.

Proposition 8

MLE provides the sample mean

$$\mu_{ML} := \frac{1}{N}(\mathbf{x}^1 + \cdots + \mathbf{x}^N), \tag{5.4}$$

as estimate for μ and the discrete sample variance with respect to the sample mean

$$\sigma^2_{ML} = \frac{1}{dN} \sum_{j=1}^{N} \|\mathbf{x}_j - \mu_{ML}\|_2^2 \tag{5.5}$$

as an estimate for σ^2.

Proof of Proposition 8: Maximizing $p(\mathbf{x}^1, \dots, \mathbf{x}^N)$ is most conveniently done by maximizing its logarithm

$$\log p(\mathbf{x}^1, \dots, \mathbf{x}^N) = -\frac{1}{2\sigma^2} \sum_{j=1}^{N} \|\mathbf{x}_j - \mu\|_2^2 - \frac{dN}{2} \log(2\sigma^2) - \frac{dN}{2} \log(\pi) \quad \text{(log-likelihood function)}. \quad (5.6)$$

Maximization over μ is independent of σ^2. Taking $E(\mu) := \sum_{j=1}^{N} \|\mathbf{x}_j - \mu\|_2^2$, one has $\nabla E(\mu) = 2\sum_{j=1}^{N}(\mathbf{x}^j - \mu) = 0 \;\Leftrightarrow\; \mu = \mu_{ML}$.

Take $a := (2\sigma^2)^{-1}$, it suffices to maximize over a. Differentiation with respect to a and setting the derivative to zero, yields the unique solution a_{ML} by

$$0 = -\sum_{j=1}^{N} \|\mathbf{x}_j - \mu_N\|_2^2 + \frac{dN}{2}\frac{1}{a_{ML}} \quad \Rightarrow \quad 2\sigma_{ML}^2 = \frac{1}{a_{ML}} = \frac{2}{dN} \sum_{j=1}^{N} \|\mathbf{x}_j - \mu_N\|_2^2$$

which is (5.5) $\square$

Remark 9

The estimates μ_{ML}, σ_{ML}^2 are independent of wether the data are sampled according to $\mathcal{N}(\cdot|\mu, \sigma^2\mathbf{I})$ or $w\mathcal{N}(\cdot|\mu, \sigma^2\mathbf{I})$ where $w > 0$ is any "weight factor". How to determine such a weight?

Maximum Likelihood Estimator (MLE)

Remark 10

This can be generalized to non-spherical Gaussians $\mathbf{X} \sim \mathcal{N}(\mu; \boldsymbol{A})$, *i.e.,*

$$\mathcal{N}(\mathbf{x}|\mu, \boldsymbol{A}) := \frac{1}{(2\pi)^{d/2}|\det \boldsymbol{A}|^{1/2}} \exp\Big\{ -\frac{1}{2}(\mathbf{x}-\mu)^\top \boldsymbol{A}^{-1}(\mathbf{x}-\mu).$$

One obtains $\mu_{ML} = \frac{1}{N}\sum_{j=1}^N \mathbf{x}^j$ *as before and*

$$\boldsymbol{A}_{ML} = \frac{1}{N}\sum_{j=1}^N (\mathbf{x}^j - \mu_{ML})(\mathbf{x}^j - \mu_{ML})^\top.$$

Hint: the joint density of $\mathbf{X}^1, \dots, \mathbf{X}^N \sim \mathcal{N}(\mu; \boldsymbol{A})$ is (by independence)

$$p(\mathbf{x}^1, \dots, \mathbf{x}^N) = \prod_{j=1}^N \mathcal{N}(\mathbf{x}^j|\mu, \boldsymbol{A}) = \frac{1}{(2\pi)^{dN/2}|\det \boldsymbol{A}|^{N/2}} e^{-\frac{1}{2}\sum_{j=1}^N (\mathbf{x}^j-\mu)\boldsymbol{A}^{-1}(\mathbf{x}^j-\mu)} \rightsquigarrow$$

maximize over μ and $\boldsymbol{R} = \boldsymbol{A}^{-1}$

$$0 \stackrel{!}{=} \log p(\mathbf{x}^1, \dots, \mathbf{x}^N) = -\frac{dN}{2}\log(2\pi) - \frac{N}{2}\log|\det \boldsymbol{A}| - \frac{1}{2}\sum_{j=1}^N (\mathbf{x}^j - \mu)\boldsymbol{A}^{-1}(\mathbf{x}^j - \mu).$$

Maximizing over $\mu \rightsquigarrow$

$$\partial_\mu \log p(\mathbf{x}^1, \dots, \mathbf{x}^N) \stackrel{!}{=} 0 \quad \rightsquigarrow \quad 0 = \sum_{j=1}^N \boldsymbol{A}^{-1}(\mathbf{x}^j - \mu) = \boldsymbol{A}^{-1}\Big(\sum_{j=1}^N (\mathbf{x}^j - \mu)\Big) \Leftrightarrow \sum_{j=1}^N \mathbf{x}^j = N\mu.$$

Maximizing over $\boldsymbol{R} := \boldsymbol{A}^{-1} \rightsquigarrow$

$$0 \stackrel{!}{=} \frac{N}{2}\frac{d}{d\boldsymbol{R}} \log|\det \boldsymbol{R}| - \frac{1}{2}\frac{d}{d\boldsymbol{R}} \sum_{j=1}^N (\mathbf{x}^j - \mu_{ML})\boldsymbol{R}(\mathbf{x}^j - \mu_{ML})$$

Notice: (chain rule)

$$\frac{d}{d\boldsymbol{R}} \log|\det \boldsymbol{R}| = \boldsymbol{R}^{-1} = \boldsymbol{A}, \quad \frac{d}{d\boldsymbol{R}} \sum_{j=1}^N (\mathbf{x}^j - \mu)\boldsymbol{R}(\mathbf{x}^j - \mu) = \sum_{j=1}^N (\mathbf{x}^j - \mu_{ML})(\mathbf{x}^j - \mu_{ML})^\top.$$

$\rightsquigarrow$ □

How good are these estimates?

Note: for each draw $\mathbf{x}^1, \ldots, \mathbf{x}^N$ one obtains estimates $\mu_{ML} = \mu_{ML}(\mathbf{X}^1, \ldots \mathbf{X}^N)$, $\sigma_{ML} = \sigma_{ML}(\mathbf{X}^1, \ldots \mathbf{X}^N)$ which will vary over repeated draws and are therefore also random variables.

Exercise 11

μ_{ML}, σ_{ML} *are random variables distributed according to* $p(\mathbf{x}^1, \ldots, \mathbf{x}^N)$. *Hence we can compute the expectation of these quantities: show that*

$$\mathbb{E}[\mu_{ML}] = \mu, \qquad \mathbb{E}[\sigma^2_{ML}] = \left(\frac{dN-1}{dN}\right)\sigma^2. \tag{5.7}$$

Thus, the maximum likelihood estimate systematically underestimates the true variance by the factor $\frac{dN-1}{dN}$. This results from computing σ^2_{ML} based on the sample mean not the true mean.

(5.7) $\rightsquigarrow$

$$\tilde{\sigma}^2_{ML} := \frac{dN}{dN-1}\sigma^2_{ML} = \frac{1}{dN-1}\sum_{j=1}^{N}\|\mathbf{x}_j - \mu_{ML}\|_2^2$$

is an unbiased estimator. These are special effects reflecting a more general feature of maximum likelihood methods.

Gaussian Mixtures revisited

Mixture Models: form an important class of stochastic models. They have the form

$$p = w_1 p_1 + w_2 p_2 + \cdots + w_k p_k, \quad w_j \geq 0, \sum_{j=1}^{k} w_j = 1, \ p_j \text{ are known densities.} \tag{5.8}$$

The mixture weights w_i quantify the proportion of the density p_j in the whole stochastic process. Clearly, p is again a probability density.

In this section we consider the case: $p_j(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu_j, \sigma^2)$, $\mu_j, \mathbf{x} \in \mathbb{R}^d$, under the assumptions:

- d large
- $k \ll d$
- $\sigma \sim 1$

Task: Given data $\mathfrak{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^N\} \subset \mathbb{R}^d$, estimate $w_i, \mu_i, \sigma, j = 1, \dots, k$.

Recall: before $k = 2$, $\|\mu_1 - \mu_2\|_2 \geq Cd^{1/4}$; now $k > 2$ is permitted and centers are allowed to be closer to each other.

Strategy:

(i) Cluster the set of samples into k clusters $\mathcal{C}_j$, $j = 1, \dots, k$, where $\mathcal{C}_j$ corresponds to the set of samples generated according to p_j; This is based on the discussion over the next slides

(ii) determine μ_j, σ^2 for the Gaussian corresponding to the cluster $\mathcal{C}_j$, $j = 1, \dots, k$, as described in the previous section;

(iii) determine the weights by a least squares method.

(i) Is Based on: Invariance of Spherical Gaussians under Projection

Lemma 12

Let $\mathbb{U} \subset \mathbb{R}^d$ be a k-dimensional subspace. Then a spherical Gaussian density $\mathcal{N}(\mathbf{x}|\mu, \sigma^2\mathbf{I})$ restricted to $\mathbb{U}$ is (up to normalization) again a sperical Gaussian density with the same variance.

Proof: Let $\{\mathbf{u}^1, \dots, \mathbf{u}^k\} \subset \mathbb{R}^d$ be an orthonormal basis for $\mathbb{U}$. Complete the matrix $\mathbf{U}_k$ with columns $\mathbf{u}^i$, $i = 1, \dots, k$,to an orthonormal matrix $\mathbf{U} = (\mathbf{U}_k, {}_{N-k}\mathbf{U})$ for $\mathbb{R}^d$ by adding columns $\mathbf{u}^{k+1}, \dots, \mathbf{u}^N$. Then, for $\mathbf{x} = \mathbf{U}\mathbf{z} = \mathbf{U}_k\mathbf{z}' + {}_{N-k}\mathbf{U}\mathbf{z}''$, where $\mathbf{z}' = (z_1, \dots, z_k)$, $\mathbf{z}'' := (z_{k+1}, \dots, N)$,

$$\mathcal{N}(\mathbf{x}|\mu, \sigma\mathbf{I}) = \frac{1}{(\sigma^2 2\pi)^{d/2}} e^{-\frac{1}{2\sigma^2}\|\mathbf{U}(\mathbf{z}-\mathbf{U}^\top\mu)\|_2^2} = \frac{1}{(\sigma^2 2\pi)^{d/2}} e^{-\frac{1}{2\sigma^2}\|\mathbf{z}-\mathbf{U}^\top\mu\|_2^2},$$

where we have used that the Euclidean norm is invariant under orthogonal transformations. Writing $\mathbf{U}^\top\mu = (\mu', \mu'')$, noting that the restriction of $\mathbf{x}$ to $\mathbb{U}$ is $\mathbf{U}_k\mathbf{z}'$, and that $\|\mathbf{z} - \mathbf{U}^\top\mu\|_2^2 = \|\mathbf{z}' - \mu'\|_2^2 + \|\mathbf{z}'' - \mu''\|_2^2$ we get

$$\mathcal{N}(\mathbf{U}_k\mathbf{z}'|\mu, \sigma^2\mathbf{I}) = \frac{1}{(\sigma^2 2\pi)^{\frac{d-k}{2}}} e^{-\frac{1}{2\sigma^2}\|\mu''\|_2^2} \frac{1}{(\sigma^2 2\pi)^{\frac{k}{2}}} e^{-\frac{1}{2\sigma^2}\|\mathbf{z}'-\mu'\|_2^2} = C\mathcal{N}(\mathbf{z}'|\mu', \sigma^2\mathbf{I}),$$

as claimed. □

Remark 13

When $\mu \in \mathbb{U}$, i.e., $\mu = \mathbf{U}_k\mathbf{y}$, $\mathbf{y} \in \mathbb{R}^k$, one has $\mathbf{U}^\top\mu = \mathbf{U}^\top\mathbf{U}_k\mathbf{y} = \mathbf{y}$, i.e., the projected Gaussian has the same mean as the original one. Goal: find the subspace $\mathbb{U}_k$ spanned by the means of a Gaussian mixture.

Invariance of Spherical Gaussians under Projection

Remark: Perhaps a better way to understand a “projection” of a density to a subspace $\mathbb{U}$ is to see how it acts on functions that do not depend on variables orthogonal to $\mathbb{U}$. Specifically, for $\mathbf{U}, \mathbf{U}_k, {}_{N-k}\mathbf{U}, \mathbf{z}', \mathbf{z}'', \mu'\mathbf{u}''$ as above, consider any g such that
$g(\mathbf{x}) = g(\mathbf{U}\mathbf{z}) = g(\mathbf{U}_k\mathbf{z}' + {}_{N-k}\mathbf{U}\mathbf{z}'') = g(\mathbf{U}_k\mathbf{z}') =: \tilde{g}(\mathbf{z}')$

$$
\begin{aligned}
\int_{\mathbb{R}^d} g(\mathbf{x})\mathcal{N}(\mathbf{x}|\mu, \sigma^2\mathbf{I})d\mathbf{x} &= \frac{1}{(\sigma^2 2\pi)^{d/2}} \int_{\mathbb{R}^d} g(\mathbf{U}\mathbf{z}) e^{-\frac{1}{2\sigma^2}\|\mathbf{U}\mathbf{z}-\mu\|_2^2} d\mathbf{z} \quad (\text{since } |\det \mathbf{U}| = 1) \\
&= \frac{1}{(\sigma^2 2\pi)^{d/2}} \int_{\mathbb{R}^d} \tilde{g}(\mathbf{z}') e^{-\frac{1}{2\sigma^2}\|\mathbf{U}(\mathbf{z}-\mathbf{U}^\top\mu)\|_2^2} d\mathbf{z} \\
&= \frac{1}{(\sigma^2 2\pi)^{d/2}} \int_{\mathbb{R}^d} \tilde{g}(\mathbf{z}') e^{-\frac{1}{2\sigma^2}\|\mathbf{z}-\mathbf{U}^\top\mu\|_2^2} d\mathbf{z} \\
&= \underbrace{\frac{1}{(\sigma^2 2\pi)^{\frac{d-k}{2}}} \int_{\mathbb{R}^{d-k}} e^{-\frac{1}{2\sigma^2}\|\mathbf{z}''-\mu''\|_2^2} d\mathbf{z}''}_{=1} \frac{1}{(\sigma^2 2\pi)^{\frac{k}{2}}} \int_{\mathbb{R}^k} \tilde{g}(\mathbf{z}') e^{-\frac{1}{2\sigma^2}\|\mathbf{z}'-\mu'\|_2^2} d\mathbf{z}' \\
&= \frac{1}{(\sigma^2 2\pi)^{\frac{k}{2}}} \int_{\mathbb{R}^k} \tilde{g}(\mathbf{z}') e^{-\frac{1}{2\sigma^2}\|\mathbf{z}'-\mu'\|_2^2} d\mathbf{z}' \\
&= \int_{\mathbb{R}^k} \tilde{g}(\mathbf{z}')\mathcal{N}(\mathbf{z}'|\mu', \sigma^2\mathbf{I}) d\mathbf{z}'.
\end{aligned}
$$

Best-Fit Subspace to a Spherical Gaussians

Let $\mathbb{U} \subset \mathbb{R}^d$ be a k-dimensional subspace. Therefore there exists an orthonormal basis $\{\mathbf{u}^1, \ldots, \mathbf{u}^k\} \subset \mathbb{R}^d$ forming the matrix $\mathbf{U}_k$. By Lecture I, page 47, (5.26),

$$P_{\mathbb{U}}\mathbf{x} = \sum_{j=1}^{k} (\mathbf{x} \cdot \mathbf{u}^j)\mathbf{u}^j = \mathbf{U}_k \mathbf{U}_k^\top \mathbf{x} \tag{5.9}$$

is the orthogonal projection to $\mathbb{U}$.

Definition 14

Given a probability density p on $\mathbb{R}^d$. Then the subspace

$$\mathbb{U}_k := \underset{\mathbb{U} \subset \mathbb{R}^d, \dim \mathbb{U} = k}{\operatorname{argmax}} \mathbb{E}\left[\| P_{\mathbb{U}} \mathbf{X} \|_2^2 \right] \tag{5.10}$$

is called the best-fit k-dimensional subspace (w.r.t. p).

Remark 15

Intuitively, $\mathbb{U}_k = \mathbb{U}_k(p)$ is the subspace that "sees most" of the density p among all k-dimensional subspaces. Compare this with Lecture I, Theorem 42, when the density p is replaced by a point cloud forming the matrix $\boldsymbol{A}$. This subspace will be seen to contain the means of the Gaussian mixture.

Best-Fit Subspace to a Spherical Gaussians

A first central step is to identify the best-fit subspace for a mixture of k spherical Gaussians.

Theorem 16

Let the density p on $\mathbb{R}^d$ have the form (5.8) where $p_j = \mathcal{N}(\cdot|\mu_j, \sigma^2 \mathbf{I})$, $\mu_j \in \mathbb{R}^d$, $j = 1, \ldots, k$. Then the best-fit k-dimensional subspace $\mathbb{U}_k$ for this mixture contains the centers $\mu_j \in \mathbb{R}^d$, $j = 1, \ldots, k$. If the μ_j are linearly dependent, the uniquely define the subspace $\mathbb{U}_k$.

The proof is based on several lemmas.

Lemma 17

For $p = \mathcal{N}(\cdot|\mu, \sigma^2 \mathbf{I})$, $\mathbf{X} \sim \mathcal{N}(\mu; \sigma^2 \mathbf{I})$, $\mathbf{u} \in \mathbb{R}^d$, $\|\mathbf{u}\|_2 = 1$, one has

$$\mathbb{E}\left[(\mathbf{u}^\top \mathbf{X})^2\right] = \sigma^2 + (\mathbf{u}^\top \mu)^2. \tag{5.11}$$

Proof:

$$\begin{aligned}
\mathbb{E}[\|P_{\mathbb{U}_1}\mathbf{X}\|_2^2] &= \mathbb{E}[|\mathbf{u}\cdot\mathbf{X}|^2] = \mathbb{E}\left[(\mathbf{u}^\top(\mathbf{X}-\mu) + \mathbf{u}^\top\mu)^2\right] \\
&= \mathbb{E}\left[(\mathbf{u}^\top(\mathbf{X}-\mu))^2 + 2(\mathbf{u}^\top\mu)(\mathbf{u}^\top(\mathbf{X}-\mu)) + (\mathbf{u}^\top\mu)^2\right] \\
&= \mathbb{E}\left[(\mathbf{u}^\top(\mathbf{X}-\mu))^2\right] + 2(\mathbf{u}^\top\mu)\mathbf{u}^\top\mathbb{E}\left[\mathbf{X}-\mu\right] + (\mathbf{u}^\top\mu)^2 \\
&= \mathbb{E}\left[(\mathbf{u}^\top(\mathbf{X}-\mu))^2\right] + (\mathbf{u}^\top\mu)^2 = \sum_{j=1}^d u_j^2\mathbb{E}[(X_j-\mu_j)^2] = \sigma^2 + (\mathbf{u}^\top\mu)^2.
\end{aligned}$$

Best-Fit Subspace to a Spherical Gaussians

Lemma 18

For $p = \mathcal{N}(\cdot|\mu, \sigma^2 \mathbf{I})$ a k-dimensional subspace is a best-fit subspace for p if and only if it contains μ.

Proof: For $\mu = \mathbf{0}$, by symmetry, every k-dimensional subspace is a best-fit subspace. Assume now $\mu \neq \mathbf{0}$.

For $k = 1$, $\mathbb{U} = \mathrm{span}\,\{\mathbf{u}\}$, one has $P_{\mathbb{U}}\mathbf{x} = (\mathbf{u}\cdot\mathbf{x})\mathbf{u}$ and hence $\|P_{\mathbb{U}}\mathbf{X}\|_2^2 = (\mathbf{u}\cdot\mathbf{X})^2$. In view of (5.11), $\mathbb{E}\left[\|P_{\mathbb{U}}\mathbf{X}\|_2^2\right]$ is maximized if and only if $\mathbf{u}$ is parallel to μ, i.e., $|\mathbf{u}\cdot\mu| = \|\mathbf{u}\|_2\|\mu\|_2 = \|\mu\|_2 \rightsquigarrow \mu \in \mathbb{U}$.

For $k > 1$: suppose $\mu \notin \mathbb{U}$. Since the orthogonal complement $\mu^\perp$ of μ in $\mathbb{R}^d$ has dimension $d-1$ and $\mathbb{U}$ has dimension k we must have $\dim(\mathbb{U}\cap\mu^\perp) = k-1$. Therefore, there exists an orthonormal basis $\{\mathbf{u}^1, \dots, \mathbf{u}^{k-1}, \mathbf{u}^k\}$ of $\mathbb{U}$ where

$$\mu^\top \mathbf{u}^j = 0, \quad j = 1, \dots, k-1. \tag{5.12}$$

As before, denoting by $\mathbf{U}_r$ the matrices with columns $\mathbf{u}^1, \dots, \mathbf{u}^r$, we recall from (5.9) that $P_{\mathbb{U}}\mathbf{x} = \mathbf{U}_k\mathbf{U}_k^\top\mathbf{x}$ and (since $\mathbf{U}_k^\top\mathbf{U}_k = \mathbf{I}_k$)

$$\|P_{\mathbb{U}}\mathbf{x}\|_2^2 = (P_{\mathbb{U}}\mathbf{x})^\top P_{\mathbb{U}}\mathbf{x} = \mathbf{x}^\top\mathbf{U}_k\mathbf{U}_k^\top\mathbf{U}_k\mathbf{U}_k^\top\mathbf{x} = \mathbf{x}^\top\mathbf{U}_k\mathbf{U}_k^\top\mathbf{x} = \sum_{j=1}^k (\mathbf{x}^\top\mathbf{u}^j)^2 \quad \rightsquigarrow \text{consider} \tag{5.13}$$

$\mathbb{E}\left[\|P_{\mathbb{U}}\mathbf{X}\|_2^2\right] = \sum_{j=1}^k \mathbb{E}\left[(\mathbf{X}^\top\mathbf{u}^j)^2\right] \overset{(5.11),(5.12)}{=} (k-1)\sigma^2 + (\mathbf{u}^k\cdot\mu)^2$ maximal iff $\mathbf{u}^k = a\mu$. □

Best-Fit Subspace to a Spherical Gaussians

Proof of Theorem 16: Let $p = w_1 p_1 + \cdots + w_k p_k$ be the Gaussian mixture (i.e., $p_j(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu_j, \sigma^2 \mathbf{I})$) and let $\mathbb{U}$ be any subspace of $\mathbb{R}^d$ of dimension k. It can be spanned by an orthonormal basis $\{\mathbf{u}^1, \ldots, \mathbf{u}^k\}$.

Then, by (5.13) and linearity of $\mathbb{E}$,

$$\mathbb{E}_{\sim p}\left[\|P_{\mathbb{U}}\mathbf{X}\|_2^2\right] = \sum_{l=1}^{k} w_l \mathbb{E}_{\sim p_l}\left[\|P_{\mathbb{U}}\mathbf{X}\|_2^2\right].$$

This sum is maximized if each summand is maximized. By Lemma 18, this is the case if and only if $\mathbb{U}$ contains the means $\mu_j, j = 1, \ldots, k$. □

Outline of a Separation Algorithm

1. (Ideally) find the best-fit subspace $\mathbb{U}_k$ that contains the centers μ_j, $j = 1, \dots, k$.
2. By Lemma 12, the projection of a spherical Gaussian to $\mathbb{U}_k$ is still (now a k-dimensional) Gaussian with the same variance σ^2.
3. Suppose $\mathfrak{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^N\} \subset \mathbb{R}^d$ is the given set of samples from the mixture distribution. Let $\mathfrak{X}^k = \{\mathbf{x}^{1,k}, \dots, \mathbf{x}^{N,k}\}| \subset \mathbb{U}_k$ be the projected sample set, i.e., $\mathbf{x}^{j,k} = P_{\mathbb{U}_k}\mathbf{x}^j$, $j = 1, \dots, k$, and denote by $\Delta_{i,j} := \|\mathbf{x}^{j,k} - \mathbf{x}^{i,k}\|_2$ the mutual distances in $\mathbb{U}_k$.
   Note: since the centers μ_j already belong to $\mathbb{U}_k$ their distances don't change under projection
   $$\|\mu_j - \mu_i\|_2 = \|P_{\mathbb{U}_k}(\mu_j - \mu_i)\|_2, \quad i \neq j \leq k. \tag{5.14}$$
4. By the methods discussed in the preceding section, one can separate Gaussians in $\mathbb{R}^k$ provided that their centers satisfy
   $$\|\mu_i - \mu_j\|_2 \geq Ck^{1/4}, \tag{5.15}$$
   which is only a small threshold (independent of d) when k is bounded uniformly.
5. Exploit the latter fact to cluster $\mathfrak{X}^k$ into k clusters $\mathcal{C}_j$, $j = 1, \dots, k$, where now with high probability the points in $\mathcal{C}_j$ come from the Gaussian $p_j = \mathcal{N}(\cdot|\mu_j, \sigma^2\mathbf{I})$.
6. Compute for each $\mathcal{C}_j$ estimates $\mu_{j,ML}, \sigma^2_{j,ML}$ by means of the Maximum-Likelihood Estimator (in $\mathbb{R}^k$, see Remark 13) from the previous section, and set $\sigma^2 = \frac{1}{k}\sum_{j=1}^k \sigma^2_{j,ML}$.
7. Set $\mathbf{M} := (\mu_{1,ML}, \dots, \mu_{k,ML}) \in \mathbb{R}^{d\times k}$, $\mathbf{y} := \frac{1}{N}\sum_{j=1}^N \mathbf{x}^j$, $\rightsquigarrow \mathbf{y} \approx \mathbb{E}_{\sim p}[\mathbf{X}] = \sum_{l=1}^k w_l\mu_l$, $\rightsquigarrow \mathbf{y} \approx \sum_{l=1}^k w_l\mu_{l,ML}$; compute $\mathbf{w} = (w_1, \dots, w_k)^\top \in \mathbb{R}^k$ by $\mathbf{w} = \mathrm{argmin}_{\mathbf{v}\geq 0}\|\mathbf{Mv} - \mathbf{y}\|_2^2$.

Outline of a Separation Algorithm

Items (1) and (5) in the above sketch require further comments:

ad (1): One cannot compute the exact best-fit subspace $\mathbb{U}_k$ because one cannot carry out the required maximization exactly.

Simple idea: maximize instead with respect to the empirical mean, i.e.,

$$\operatorname*{argmax}_{\dim \mathbb{U}=k} \mathbb{E}\left[\|P_{\mathbb{U}}\mathbf{X}\|_2^2\right] \;\leftrightarrow\; \operatorname*{argmax}_{\dim \mathbb{U}=k} \left\{\frac{1}{N}\sum_{i=1}^{N} \|P_{\mathbb{U}}\mathbf{x}^i\|_2^2\right\} \tag{5.16}$$

Consider first $k=1$, $\mathbb{U} = \operatorname{span}\{\mathbf{u}\}$, $\|\mathbf{u}\|_2 = 1$, $\rightsquigarrow$

$$\mathbf{u}^1 = \operatorname*{argmax}_{\|\mathbf{u}\|_2=1} \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}^i\cdot\mathbf{u})^2. \tag{5.17}$$

Let $\boldsymbol{A}$ denote the matrix whose rows are the $\mathbf{x}^i$, i.e., $\boldsymbol{A} \in \mathbb{R}^{N\times d}$. Then, (5.17) can be equivalently restated as

$$\mathbf{u}^1 = \mathbf{u}^1(\mathfrak{X}) = \operatorname*{argmax}_{\|\mathbf{u}\|_2=1} \|\boldsymbol{A}^\top\mathbf{u}\|_2^2 = \operatorname*{argmax}_{\|\mathbf{u}\|_2=1} \mathbf{u}^\top\boldsymbol{A}\boldsymbol{A}^\top\mathbf{u}. \tag{5.18}$$

As shown in Lecture I (see e.g. the proof ofTheorem 39, or Lemma 43), $\mathbf{u}^1$ is the first left singular vector of the matrix $\boldsymbol{A}$ and

$$\max_{\|\mathbf{u}\|_2=1} \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}^i\cdot\mathbf{u})^2 = \frac{\sigma_{1,\mathfrak{X}}^2}{N}, \quad \text{(where } \sigma_{1,\mathfrak{X}} \text{ is the largest singular value of } \boldsymbol{A}.) \tag{5.19}$$

Outline of a Separation Algorithm

Returning to (5.16), we take up on the PCA Greedy Construction of the SVD in Lecture I, page 75, (6.16) and successively maximize at the ith stage $\mathbf{u}^\top \boldsymbol{A}\boldsymbol{A}^\top \mathbf{u}$ over those unit vectors $\mathbf{u}$, $\|\mathbf{u}\|_2 = 1$, which are orthogonal to the previously computed directions $\mathbf{u}^1, \dots, \mathbf{u}^{i-1}$ for $i \le k$. Hence, for $r \le k$

$$\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}^i \cdot \mathbf{u}^r)^2 = \max_{\|\mathbf{u}\|_2=1;\mathbf{u}\perp\mathbf{u}^s, s<r} \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}^i \cdot \mathbf{u})^2. \tag{5.20}$$

Let us again denote by $\mathbf{U}_k$ the matrix whose columns are these pairwise orthonormal vectors $\mathbf{u}^i$. Thus $P_{\mathbb{U}_k}\mathbf{x} = \mathbf{U}_k\mathbf{U}_k^\top \mathbf{x}$ and

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\|P_{\mathbb{U}_k}\mathbf{x}^i\|_2^2 &= \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}^i)^\top \mathbf{U}_k\mathbf{U}_k^\top \mathbf{U}_k\mathbf{U}_k^\top \mathbf{x}^i = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}^i)^\top \mathbf{U}_k\mathbf{U}_k^\top \mathbf{x}^i. = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k}(\mathbf{u}^j \cdot \mathbf{x}^i)^2 \\
&= \sum_{j=1}^{k}\left\{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{u}^j \cdot \mathbf{x}^i)^2\right\} = \sum_{j=1}^{k}\frac{\sigma_{j,\mathfrak{X}}^2}{N},
\end{aligned}$$

i.e., in view of (5.20), each summand in the curly brackets is maximized by the greedy basis.

Outline of a Separation Algorithm

Corollary 19

Step (1) in the algorithm can be realized approximately by computing the SVD of the point cloud $\boldsymbol{A}^\top \leftrightarrow \mathfrak{X}$. The subspace generated by the first left singular vectors $\mathbf{u}^i$, $i = 1, \ldots, k$, is an approximation to the exact best-fit subspace The larger the number N of samples $\mathbf{x}^i$, the closer is the empirical mean to the true expectation, i.e., the discrete maximization in (5.16) *yields better and better approximations to the exact best-fit subspace. The singuar values $\sigma^2_{j,\mathfrak{X}}$ are approximations of σ^2.*

The accuracy of the SVD based subspace affects the accuracy of the estimation for the means $\mu_{j,ML}$ taking place in the approximate subspace.

ad (5):

- Compute first all pairwise distances $\Delta_{i,j}$ (in $\mathbb{U}_k$) and order them by increasing size Δ_{i_r,j_r}; pick the smallest $r = s$ such that $\Delta_{i_s,j_s} \geq \sqrt{2d} + a =: \delta$; find a, c such that $\Delta_{i_s,j_s} \geq \sqrt{2d} + ck^{1/4} =: \Delta$ holds for all $s > r$.
- Put all pairs (i,j) into $\mathcal{S}$, for which $\|\mathbf{x}^{j,k} - \mathbf{x}^{i,k}\|_2 \leq \delta$, put all pairs with $\|\mathbf{x}^{j,k} - \mathbf{x}^{i,k}\|_2 \geq \Delta$ into $\mathcal{L}$.

Outline of a Separation Algorithm

- Consider the triangular array

$$T = \begin{pmatrix} (1,2), & (1,3), & (1,4), & \dots & ,(1,N) \\ & (2,3), & (2,4), & \dots & ,(2,N) \\ & & \vdots & \dots & ,\vdots \\ & & & & ,(N-1,N) \end{pmatrix}$$

  Let $T_{\mathcal{S}}$ be the sub-array for which all pairs belong to $\mathcal{S}$. Two pairs are connected if the have a common index. A subset of pairs is connected if any two of them can be connected by a path of connected pairs. The "content" of a connected subset is the set of involved indices. Each cluster $\mathcal{C}_j$ corresponds to the content of a maximal connected set of pairs in $T_{\mathcal{S}}$.
- Exercise: Design an efficient way of finding these sets.