High-dimensional Gaussians

Daniel Hsu

COMS 4772

Gaussian distributions

Gaussian (normal) distributions

- $Z \sim \mathrm{N}(0,1)$ means Z follows a *standard Gaussian distribution*, i.e., has probability density

$$z \;\mapsto\; \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \,.$$

- If $Z_1, Z_2, \dots, Z_d$ are iid $\mathrm{N}(0,1)$ random variables, then say $\boldsymbol{Z} := (Z_1, Z_2, \dots, Z_d)$ follows a *standard multivariate Gaussian distribution* on $\mathbb{R}^d$, i.e., $\boldsymbol{Z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$.
- Other Gaussian distributions on $\mathbb{R}^d$ arise by applying *(invertible) linear maps* and *translations* to $\boldsymbol{Z}$:

$$\overbrace{\boldsymbol{z} \;\mapsto\; \boldsymbol{A}\boldsymbol{z}}^{\text{linear map}} \;\mapsto\; \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu} \,. \qquad \underbrace{\boldsymbol{A}\boldsymbol{z} \;\mapsto\; \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu}}_{\text{translation}}$$

  - $\boldsymbol{X} := \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{A}^\top)$ has

$$\mathbb{E}(\boldsymbol{X}) \;=\; \boldsymbol{\mu} \quad \text{and} \quad \operatorname{cov}(\boldsymbol{X}) \;=\; \boldsymbol{A}\boldsymbol{A}^\top \,.$$

Shape of Gaussian distributions

- Let $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu} \in \mathbb{R}^d$, and $\boldsymbol{\Sigma} \succ \boldsymbol{0}$.
- Contours of equal density are ellipsoids around $\boldsymbol{\mu}$:

$$\{\boldsymbol{x} \in \mathbb{R}^d : (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = r^2\} \,.$$

- Let eigenvalues of $\boldsymbol{\Sigma}$ be $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$, corresponding (orthonormal) eigenvectors be $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_d$.
  - $\operatorname{var}(\langle \boldsymbol{v}_i, \boldsymbol{X} \rangle) = \lambda_i$. (This is true even if $\boldsymbol{X}$ is not Gaussian.)
  - If $Y_i := \langle \boldsymbol{v}_i, \boldsymbol{X} - \boldsymbol{\mu} \rangle$, then $Y_i \sim \mathrm{N}(0, \lambda_i)$.
  - $Y_1, Y_2, \dots, Y_d$ are independent; $\boldsymbol{Y} := (Y_1, Y_2, \dots, Y_d) \sim \mathrm{N}(\boldsymbol{0}, \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_d))$.
- What about concentration properties?

Concentration of spherical Gaussians

- Spherical Gaussian: $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$.
- Pick any $\delta \in (0,1)$. Then

$$\text{for any } \boldsymbol{u} \in S^{d-1}, \quad \mathbb{P}\left(\langle \boldsymbol{u}, \boldsymbol{X} - \boldsymbol{\mu} \rangle \leq \sigma \sqrt{2 \ln(1/\delta)}\right) \;\geq\; 1 - \delta \,,$$

$$\mathbb{P}\left(\|\boldsymbol{X} - \boldsymbol{\mu}\|_2^2 \leq \sigma^2 d \left(1 + 2\sqrt{\frac{\ln(1/\delta)}{d}} + \frac{2\ln(1/\delta)}{d}\right)\right) \;\geq\; 1 - \delta \,,$$

$$\mathbb{P}\left(\|\boldsymbol{X} - \boldsymbol{\mu}\|_2^2 \geq \sigma^2 d \left(1 - 2\sqrt{\frac{\ln(1/\delta)}{d}}\right)\right) \;\geq\; 1 - \delta \,.$$

  (Standard tail bounds for $\mathrm{N}(0,1)$ and $\chi^2(d)$ distributions.)
- Behaves like spherical shell around $\boldsymbol{\mu}$ of radius $\sigma\sqrt{d}$ and thickness $O(\sigma d^{1/4})$.

Concentration of general Gaussians

- General Gaussian: $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- Concentration of $\langle \boldsymbol{u}, \boldsymbol{X} - \boldsymbol{\mu} \rangle$ for $\boldsymbol{u} \in S^{d-1}$ depends on $\boldsymbol{u}$:

$$\langle \boldsymbol{u}, \boldsymbol{X} - \boldsymbol{\mu} \rangle \;\sim\; \mathrm{N}(0, \boldsymbol{u}^\top \boldsymbol{\Sigma} \boldsymbol{u}) \,.$$

- Concentration of $\|\boldsymbol{X} - \boldsymbol{\mu}\|_2^2$: *a mismatch of norms.*
  - $\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})\|_2^2 \sim \chi^2(d)$.
  - $\|\boldsymbol{X} - \boldsymbol{\mu}\|_2^2$ distributed as linear combination of independent $\chi^2(1)$ random variables:

$$\sum_{i=1}^{d} \lambda_i Z_i^2$$

    where $Z_1, Z_2, \ldots, Z_d$ are iid $\mathrm{N}(0,1)$.
  - $\mathbb{E}\,\|\boldsymbol{X} - \boldsymbol{\mu}\|_2^2 = \sum_{i=1}^{d} \lambda_i$.
  - $\|\boldsymbol{X} - \boldsymbol{\mu}\|_2^2$ is $(4\sum_{i=1}^{d} \lambda_i^2, 4\lambda_1)$-subexponential.

Eccentricity of general Gaussians

- For $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with probability $1 - \delta$,

$$\|\boldsymbol{X} - \boldsymbol{\mu}\|_2^2 \;\in\; \bar{\lambda} d \left(1 \pm O\left(\sqrt{\frac{\kappa \log(1/\delta)}{d}} + \frac{\kappa \log(1/\delta)}{d} \right) \right),$$

  where $\bar{\lambda} := \frac{1}{d}\sum_{i=1}^{d} \lambda_i$ and $\kappa := \lambda_1 / \bar{\lambda}$.
- κ measure *eccentricity* of equal density ellipsoids: $1 \leq \kappa \leq d$.

Using multivariate Gaussians as a statistical model

- $\mathcal{P} := \{\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) : \boldsymbol{\mu} \in \mathbb{R}^d,\ \boldsymbol{\Sigma} \succ \mathbf{0}\}$
- Parameter estimation given data $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$
- Maximum likelihood estimators:

$$\hat{\boldsymbol{\mu}} \;:=\; \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\,, \qquad \widehat{\boldsymbol{\Sigma}} \;:=\; \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^{\top}\,.$$

- Accuracy when data is iid sample from $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \;\leq\; ?\,, \qquad \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_? \;\leq\; ?$$

  - $\hat{\boldsymbol{\mu}} \;\sim\; \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$.
  - $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2 \;=\; \max_{\boldsymbol{u} \in S^{d-1}} \left| \boldsymbol{u}^{\top} (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \boldsymbol{u} \right|$.
  - Note that $\mathbb{E}(\widehat{\boldsymbol{\Sigma}}) \neq \boldsymbol{\Sigma}$, but almost.

Multiple Gaussian populations

Multiple populations

- Often data do not come from just a single population, but rather **several different populations**.
- **If data are "labeled" by population**, then can partition data, and (say) fit a Gaussian distribution to each part (or whatever).
- **What if data are not labeled?**

Simple case: multiple Gaussian populations

- Suppose data come from k populations $P_1, P_2, \ldots, P_k$.
- Further, for extreme simplicity, suppose $P_i = \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{I})$.
- When can we separate data from P_i and P_j ($i \neq j$)?
  - Easier when means $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ are farther apart.
- **Strict separation condition**:
  - Whenever $\boldsymbol{a}$ and $\boldsymbol{b}$ come from same P_i, and $\boldsymbol{c}$ comes from different P_j,
$$\|\boldsymbol{a} - \boldsymbol{b}\|_2 \; < \; \|\boldsymbol{a} - \boldsymbol{c}\|_2 \, .$$
- Under strict separation, Kruskal's minimum spanning tree (where edge weight = Euclidean distance) *connects data from same population, before connecting across populations*.
- How far apart should $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ be to have strict separation?

Disjoint spherical shells

- Recall: $\mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{I}) \approx$ thin spherical shell around $\boldsymbol{\mu}_i$ of radius $\sqrt{d}$.
- If $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2 \gg \sqrt{d}$, then "$\mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{I}) \cap \mathrm{N}(\boldsymbol{\mu}_j, \boldsymbol{I}) \approx 0$".
  - (This can be easily formalized.)
- But this reasoning ignores approximate orthogonality!

Approximate orthogonality

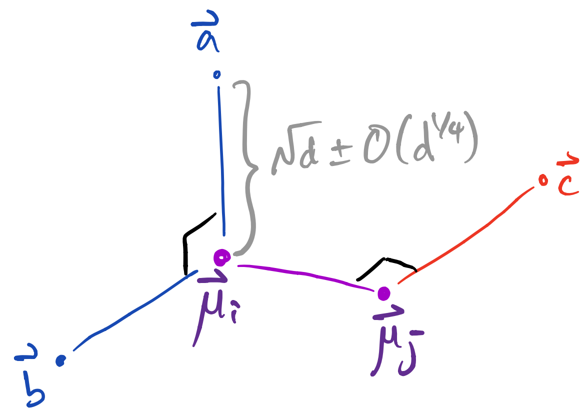

Figure 1: Distances between points from spherical Gaussian populations

Probabilistic analysis

- Let $\boldsymbol{A}, \boldsymbol{B} \sim \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{I})$ and $\boldsymbol{C} \sim \mathrm{N}(\boldsymbol{\mu}_j, \boldsymbol{I})$ (all independent).
- Write

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{\mu}_i + \boldsymbol{Z}_A\,, & \boldsymbol{C} &= \boldsymbol{\mu}_j + \boldsymbol{Z}_C\,, \\ \boldsymbol{B} &= \boldsymbol{\mu}_i + \boldsymbol{Z}_B\,, \end{aligned}$$

  where $\boldsymbol{Z}_A, \boldsymbol{Z}_B, \boldsymbol{Z}_C$ are iid $\mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$.
- Then

$$\begin{aligned} &\|\boldsymbol{A} - \boldsymbol{C}\|_2^2 - \|\boldsymbol{A} - \boldsymbol{B}\|_2^2 \\ &\quad = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2 + 2\langle \boldsymbol{\mu}_i - \boldsymbol{\mu}_j, \boldsymbol{Z}_A - \boldsymbol{Z}_C \rangle + \|\boldsymbol{Z}_A - \boldsymbol{Z}_C\|_2^2 \\ &\qquad - \|\boldsymbol{Z}_A - \boldsymbol{Z}_B\|_2^2\,. \end{aligned}$$

- With high probability, this is at least

$$\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2 - O(\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2) - O(\sqrt{d})\,,$$

  which is positive when $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2 \gg d^{1/4}$.

Probabilistic analysis (continued)

- Need previous concentration to hold for all triples in n data: union bound over $O(n^3)$ events means we need $\log(n)$ factors in separation, specifically

$$\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2 \;\geq\; C\Big((d\log(n))^{1/4} + \log(n)\Big) \quad \text{for all } i \neq j\,,$$

  where $C > 0$ is a sufficiently large absolute constant.

Mixture models

- Can think of overall population as a *mixture distribution*

$$\pi_1\,\mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{I}) + \pi_2\,\mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{I}) + \cdots + \pi_k\,\mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{I})\,,$$

  where π_i is expected proportion from $\mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{I})$.
- Usually MLE for mixture distribution parameters $\{(\pi_i, \boldsymbol{\mu}_i)\}_{i=1}^k$ is computationally intractable in general.
- But with **strict separation**:
  - First separate data by *mixture component* source.
  - Then estimate π_i and $\boldsymbol{\mu}_i$ using separated data.

Another approach

- Project data to line spanned by some $\boldsymbol{u} \in S^{d-1}$.
- With "good" $\boldsymbol{u}$, projected means remain separated.
  - Use classical statistical methods to estimate projected means.
- Do this for d nearby but linearly independent $\boldsymbol{u}$; can then back-out estimates of original means.

Projection pursuit

Exploratory data analysis (Tukey)

- Many classical data analysis methods based on finding "interesting" features of data set.
- E.g., visually inspect many one-dimensional projections of data.
- Called **projection pursuit**.
- Folklore: most projections are not interesting.

Two examples

- $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)$ is Rademacher (i.e., uniform on $\{\pm 1\}^d$).
- $\boldsymbol{u}_1 := (1/\sqrt{d}, 1/\sqrt{d}, \ldots, 1/\sqrt{d})$:

$$\langle \boldsymbol{u}_1, \boldsymbol{X} \rangle \;=\; \frac{1}{\sqrt{d}} \sum_{i=1}^{d} X_i \,.$$

  - By central limit theorem, this is approximately $\mathrm{N}(0, 1)$.
- $\boldsymbol{u}_2 := (1, 0, \ldots, 0)$:

$$\langle \boldsymbol{u}_2, \boldsymbol{X} \rangle \;=\; X_1 \,.$$

  - Very different from $\mathrm{N}(0, 1)$.
- **"Theorem"**: Most projections are more like $\boldsymbol{u}_1$ rather than $\boldsymbol{u}_2$.

Projection pursuit asymptotics (Diaconis-Freedman, 1984)

- Suppose $X_1, X_2, \ldots, X_d$ are independent random variables.
  - Assume $\mathbb{E}(X_i) = 0$, $\mathbb{E}(X_i^2) = 1$, $\mathbb{E}\,|X_i|^3 \le \rho < \infty$.
- For nearly all $\boldsymbol{u} \in S^{d-1}$,

$$\sup_{t \in \mathbb{R}} \Big| \mathbb{P}(\langle \boldsymbol{u}, \boldsymbol{X} \rangle \le t) - \Phi(t) \Big| \; \le \; \tilde{O}\left(\frac{\rho}{\sqrt{d}} \right),$$

where Φ is $\mathrm{N}(0,1)$ CDF.

Application to mixture models

- Suppose $\boldsymbol{X} \sim \pi_1 P_1 + \pi_2 P_2 + \cdots + \pi_k P_k$, where each P_i is a *product distribution*.
- $\boldsymbol{X}$ generally does not have independent coordinates.
- But for most $\boldsymbol{u} \in S^{d-1}$, $\langle \boldsymbol{u}, \boldsymbol{X} \rangle$ is close to

$$\pi_1 N_1 + \pi_2 N_2 + \cdots + \pi_k N_k$$

for some univariate normal distributions $N_1, N_2, \ldots, N_k$.