# Lecture 2 : MLE, Gaussian Models, PCA

## 1 Maximum likelihood estimation (MLE)

**Setup:** Distribution $p(x;\theta)$ parameterized by $\theta \in \Theta$

I.I.D. observations $\{x^{(i)}\}_{i=1..n}$

**Goal:** Learn the parameters $\theta$

**Def:** The maximum likelihood estimator is given by

$$\hat{\theta}^{MLE} := \underset{\theta \in \Theta}{\operatorname{argmax}} \left\{ \mathcal{L}(\theta) := \prod_{i=1}^{n} p(x^{(i)};\theta) \right\}$$

**Examples:**

- **Naive Bayes**

$$y \sim \text{Ber}(\pi)$$
$$x_k \mid y \sim \text{Ber}(\theta_{yk}) \quad , k = 1, \dots, K$$

e.g.: $x = (x_1, \dots, x_K)$ document in Bag-of-Words format

$y \in \{0,1\}$ : spam vs non-spam

Observe dataset $\{(x^{(i)}, y^{(i)})\}_{i=1..n}$

log-likelihood:

$$\ell(\pi, \theta) = \log \prod_{i=1}^{n} p(y^{(i)};\pi) \prod_{k=1}^{K} p(x_k^{(i)} \mid y^{(i)};\theta)$$
$$= \sum_{i=1}^{n} y^{(i)} \log \pi + (1-y^{(i)}) \log(1-\pi)$$
$$+ \sum_{i} \sum_{k} x_k^{(i)} \log \theta_{y^{(i)}k} + (1-x_k^{(i)}) \log(1-\theta_{y^{(i)}k})$$

- $\partial_\pi \ell(\pi, \theta) = \frac{\sum_{i=1}^{m} y^{(i)}}{\pi} - \frac{\sum_i (1-y^{(i)})}{1-\pi} = 0$

$$(1-\pi)\sum_i y^{(i)} - \pi \sum_i (1-y^{(i)}) = 0$$

$$\sum_i y^{(i)} - \pi \cdot m = 0$$

$$\Rightarrow \boxed{\hat{\pi} = \frac{\sum_i y^{(i)}}{m}}$$

- $\partial_{\theta_{yk}} \ell(\pi, \theta) = 0$ (similar)

$$\Rightarrow \boxed{\hat{\theta}_{yk} = \frac{\sum_i 1\{y^{(i)} = y\}\, x_k^{(i)}}{\sum_i 1\{y^{(i)} = y\}}}$$

**Remark:** (Generative vs Discriminative models)

→ In Naive Bayes, we model the JOINT distribution $p(x, y)$ and predict using $p(y|x) \propto p(x|y) \cdot p(y)$

This is known as **generative / joint modeling**

→ Often in ML, we might only care about **predictions** using $p(y|x)$.

**Discriminative / conditional** learning directly estimates a model of $p(y|x;\theta)$ using "conditional" likelihood:

$$\hat{\theta} = \arg\max_\theta \prod_i p(y^{(i)}|x^{(i)};\theta)$$

→ avoids difficulties of modeling $p(x)$

Ex: . Linear regression $y|x \sim \mathcal{N}(\theta^T x, \sigma^2 I)$

. Logistic regression $y|x \sim \text{Ber}(\sigma(\theta^T x))$

$$\sigma(u) = \frac{e^u}{e^u + 1} \quad \text{(sigmoid)}$$

# 2. Gaussian models & PCA

**Setup:** Data analysis with a sample $\{x_i\}_{i=1..n}$, $x_i \in \mathbb{R}^d$ i.i.d. from some distribution $p(x)$.

**Key quantities:**

→ First moment / mean

$$\mu = \mathbb{E}_p x, \qquad \hat{\mu} = \mathbb{E}_{\hat{p}} x = \frac{1}{n}\sum_i x_i$$

→ Second moment $\mathbb{E}\, x x^T$

Covariance

$$\Sigma = \mathbb{E}\left[(x-\mu)(x-\mu)^T\right], \qquad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})^T$$

- $\mu$ ↔ "center of gravity" $\mu = \arg\min_v \mathbb{E}\|x-v\|^2$

  $\Sigma$ ↔ "spread" in each direction

# Constructing Gaussian distributions, maximum entropy

a.k.a: why are Gaussians common modeling tools?

**Def:** The **entropy** of a r.v. $X$ with distribution (density) $p(x)$ is:

$$H(X) = -\int_{\mathcal{X}} p(x) \log p(x)\, dx$$

- standard measure of "**information**", "**uncertainty**"
- originates from statistical physics, key quantity in information theory

→ "**Maximum Entropy Principle**":
pick the distribution with largest entropy,
under some moment constraints

- "non-committal"
- a form of regularity

**Ex:** for $X$ discrete r.v. on a discrete set $\mathcal{X}$ with $|\mathcal{X}| = K$,
we have $0 \leq H(X) \leq \log K$

- $H(X) = 0$ (minimal) for Dirac Delta
- $H(X) = \log K$ (maximal) for Uniform Distribution

**Prop:** (Maximum entropy for Gaussian)
The max. ent. distr. satisfying the constraints $\begin{cases} \mathbb{E}X = \mu \\ \mathbb{E}(X-\mu)(X-\mu)^T = \Sigma \end{cases}$
on the first two moments is the Gaussian $\mathcal{N}(\mu, \Sigma)$

Remarks: → Along with Central Limit Theorem, this justifies the "universality" of Gaussian distributions

→ Gaussian is thus a natural choice in many applications (e.g. signal processing)

proof: (for $d=1$)

$x \in \mathbb{R}$, find density $p(x)$ that solves the following optimization problem:

$$\max H(p)$$

$$\text{s.t.} \quad \mathbb{E}_p[X] = \mu$$
$$\mathbb{E}_p[(X-\mu)^2] = \sigma^2$$
$$\int p(x)\,dx = 1$$

→ Introduce Lagrange multipliers $\lambda, \nu, \gamma$ and Lagrangian

$$\mathcal{L}(p, \lambda, \nu) = -\int p(x) \log p(x) dx + \lambda\left(\int x p(x) dx - \mu\right)$$
$$+\nu\left(\int (x-\mu)^2 p(x) dx - \sigma^2\right) + \gamma\left(\int p(x) dx - 1\right)$$

→ Look for $p(x), \lambda, \nu$ s.t.

$$\begin{cases} \frac{\delta \mathcal{L}}{\delta (p(x))} = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} = 0 \quad (\text{i.e. } \mathbb{E}X = \mu) \\ \frac{\partial \mathcal{L}}{\partial \nu} = 0 \quad (\text{i.e. } \mathbb{E}(X-\mu)^2 = \sigma^2) \\ \frac{\partial \mathcal{L}}{\partial \gamma} = 0 \quad (\text{i.e. } \int p = 1) \end{cases}$$

$$\frac{\delta \mathcal{L}}{\delta(p(x))} = -\log p(x) - 1 + \lambda x + \nu (x-\mu)^2 + \gamma = 0$$

$$\Rightarrow \quad p(x) \propto e^{ax^2 + bx + c} \quad \text{for some } a, b, c$$

find $a, b, c$ ($\Leftrightarrow \lambda, \sigma, \mu$) s.t. constraints hold.

yields $\mathcal{N}(\mu, \sigma^2)$ ! □

Note: . First example of variational principle

. Generalizes to other moments, $\mathbb{E}_p[\phi(x)]$ (exponential families)

. Tight links with MLE through convex analysis ("moment matching")

$\mu = \hat{\mu}$, $\Sigma = \hat{\Sigma}$ } is the MLE !

---

## ■ Principal component analysis (PCA)

Q: How to represent data efficiently using second moments information?

→ PCA: find a basis of "principal components" s.t. each component captures uncorrelated "factors" with decreasing variance.

→ Variance of $\langle x, v \rangle$ is given by $\langle \Sigma v, v \rangle$

→ 1st PC: $v_1 := \underset{\|v\|=1}{\arg\max} \langle \Sigma v, v \rangle$ $\left( \langle \Sigma v_1, v_1 \rangle = \lambda_1(\Sigma) \right)$

→ 2nd PC: $v_2 := \underset{\substack{\|v\|=1 \\ v \perp v_1}}{\arg\max} \langle \Sigma v, v \rangle$ $\left( \lambda_2(\Sigma) \leq \lambda_1(\Sigma) \right)$

- **Spectral Theorem**: $\Sigma = U \Lambda U^T$

with $U U^T = U^T U = I$

$\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$

$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$

→ if $\tilde{x} := x - \mathbb{E}x$, $z := U^T \tilde{x}$ the projections of $x$ along PCs

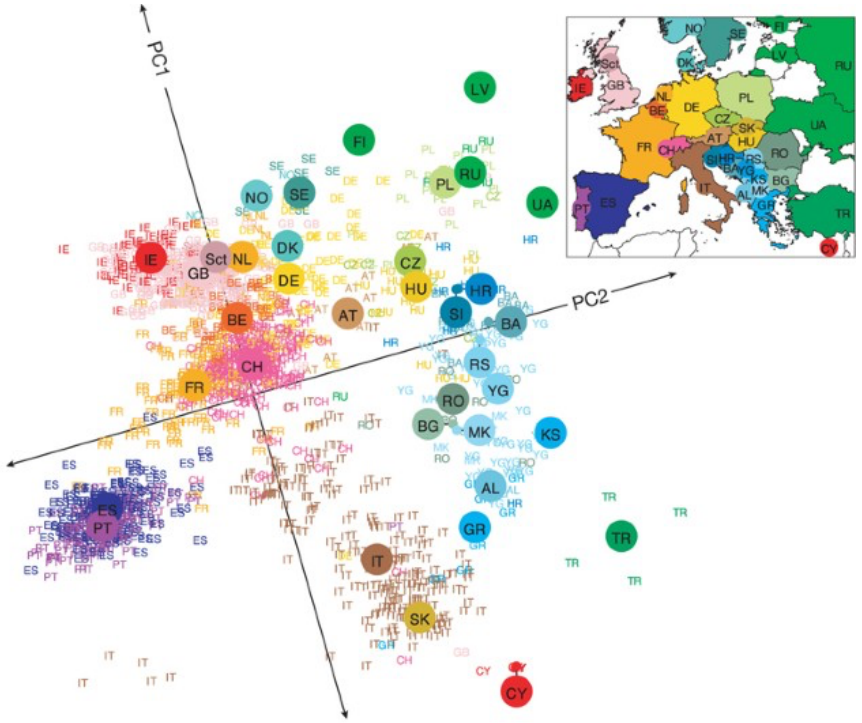
$\mathbb{E} z z^T = U^T \mathbb{E}(\tilde{x}\tilde{x}^T) U = U^T \Sigma U = \Lambda$

$\Rightarrow$ uncorrelated factors

Example:

Gene expression data (d very large)

First 2 PCs capture geography!



Notes: • An example of "latent representation" of $x$:

$$z = U^T(x-\mu) \quad \text{("analysis")}$$

$$\hat{x} = Uz + \mu \quad \text{("synthesis" / "reconstruction")}$$

• Can be viewed as "best linear approximation" in mean-squared sense

$$\min_{A \in \mathbb{R}^{d \times k}} \mathbb{E}\|x - AA^Tx\|^2$$

→ solution: $A = [u_1 \cdots u_k]$

• "Linear" representation: somewhat limited beyond Gaussian data

Real data may require representations beyond Linear:

→ multiple competing hypotheses

(e.g. which cluster in | k-means / Gaussian mixture)


→ "explaining-away" phenomenon

learning about one latent cause may help rule out others

→ requires inherently non-linear procedures

# Covariance estimation

When we only observe samples $\{x_i\}$ from $p(x)$, our models (Gaussian, or PCA analysis) crucially relies on quality of our estimate $\hat{\Sigma}$ of $\Sigma$.

**Q:** How good is $\hat{\Sigma}$ ?

- Estimate in spectral norm : $\|A\|_2 = \sup_{\|x\|=1} \|Ax\|$
- Useful for PCA : $\|\hat{\Sigma} - \Sigma\|_2$ small $\Rightarrow \lambda_i(\hat{\Sigma}) \approx \lambda_i(\Sigma)$ and $U(\hat{\Sigma}) \approx U(\Sigma)$ if $\{\lambda_i\}$ are well-separated

**Theorem** (e.g. Vershynin, HDP, Thm 4.7.1)

Assume that $\langle x, v \rangle$ is sub-gaussian for any $v \in \mathbb{R}^d$ (true e.g. if $x$ is Gaussian)

Then we have

$$\mathbb{E}\, \|\hat{\Sigma}_n - \Sigma\|_2 \leq C \cdot \left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right) \|\Sigma\|_2$$

$\hat{\Sigma}_n$ from $\{x_i\}_{i=1..n}$ i.i.d.

→ In particular, $n \sim d$ samples suffice to achieve small error

→ ND "curse of dimensionality"

⇒ Other metrics may be desirable, e.g., on estimate of the precision matrix $\Lambda = \Sigma^{-1}$, which plays an important role in structure estimation