2.Asymptotics

1. Basic question

Data: $\{x^{(1)}, \cdots, x^{(n)}\} \sim P_{\theta^*}$, where P_{θ^*} represents the unknow distribution with parameters θ^*.

Question: Can we come up with an estimate $\hat{\theta}$ that gets close to θ^*?

---

2. Gaussian mean estimation

Goal: to estimate mean of a Gassian distribution.

Suppose $\{x^{(1)}, \cdots, x^{(n)}\} \sim N(\theta^*, \sigma^2 I)$ i.i.d., where $\sigma^2 I$ is known.

Define $\hat{\theta} = \frac{1}{n} \sum x^{(i)}$, we now study $\hat{\theta} - \theta^*$.

[Lemma 1] $\hat{\theta} - \theta^* \sim N(0, \frac{\sigma^2 I}{n})$

Pf: $x^{(i)} - \theta^* \sim N(0, \sigma^2 I)$, then we have

$$S_n := \sum_{i=1}^{n} (x^{(i)} - \theta^*) \sim N(0, n\sigma^2 I)$$

since $x^{(i)} - \theta^*$ is independent with $x^{(j)} - \theta^*$ if $i \neq j$.

$$\Rightarrow \hat{\theta} - \theta^* = S_n / n \sim N(0, \frac{\sigma^2 I}{n}) \qquad \square$$

[Lemma 2] $\|\hat{\theta} - \theta^*\|_2^2 \sim \frac{\sigma^2}{n} \chi_d^2$

$E[\|\hat{\theta} - \theta^*\|_2^2] = \frac{d\sigma^2}{n}$

Pf: By Lemma 1, we have

$$v := (\hat{\theta} - \theta^*) \sqrt{\frac{n}{\sigma^2}} \sim N(0, I)$$

$$\frac{n}{\sigma^2} \|\hat{\theta} - \theta^*\|_2^2 = \sum_{j=1}^{d} v_j^2 \sim \chi_d^2$$

$\Rightarrow \|\hat{\theta} - \theta^*\|_2^2 \sim \frac{\sigma^2}{n} \chi_d^2$, which proves the first statement.

Take expectations on both sides, we have

Taking expectations on both side, we have

$$E[\|\hat{\theta}-\theta^*\|_2^2] = \frac{d\sigma^2}{n} \qquad \square$$

---

3. Multinomial estimation

Suppose we have an unknown multinomial distribution over d choises: $\theta^* \in \Delta_d$ ($\theta = [\theta_1, \cdots, \theta_d]$, $\theta_j \geq 0$ and $\Sigma \theta_j = 1$).

Suppose $\{x^{(1)}, \cdots, x^{(n)}\} \sim \text{Multinomial}(\theta^*)$, i.i.d., where $x^{(i)} \in \{e_1, \cdots, e_d\}$ and $e_j \in \{0,1\}^d$ is one-hot vector.

Consider the empirical distribution:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}.$$

Strategy: to study the asymptotic behavior of $\hat{\theta}$.

First, by Central Limit Theorem, we have

$$\sqrt{n}(\hat{\theta}-\theta^*) \xrightarrow{d} N(0, V)$$

where $V := \text{diag}(\theta^*) - \theta^*(\theta^*)^T$,

$$V_{jk} = \begin{cases} \theta_j^*(1-\theta_j^*) & , \text{ if } j=k \\ -\theta_j^*\theta_k^* & , \text{ if } j \neq k. \end{cases}$$

(Since $E[x_j] \underset{\Uparrow}{=} E[x_j^2] = \theta_j^*$, $E[x_j \underset{\Uparrow}{x_k}] = 0$ for $j \neq k$)

x_j can only be 0 or 1. at most one of them be 1.

Next, by the Continuous Mapping Theorem on $\|\cdot\|_2^2$:

$$n\|\hat{\theta}-\theta^*\|_2^2 \xrightarrow{d} tr(W(V,1)) \qquad (*)$$

where $W(V,k)$ is the Wishart distribution with mean matrix V and k degrees of freedom.

Since $z \sim N(0,V)$, then $zz^T \sim W(V,1)$, $\|z\|_2^2 = tr(zz^T)$.

Taking expectations of both sides of (*), and dividing by n:

$$E[\|\hat{\theta}-\theta^*\|_2^2] \to \left(\sum_{j=1}^{d}\theta_j^*(1-\theta_j^*)\right)\frac{1}{n} + o\left(\frac{1}{n}\right)$$
$$\le \frac{1}{n} + o\left(\frac{1}{n}\right)$$

Note: $Y_n \xrightarrow{d} Y$, if we want $E[Y_n] \to E[Y]$, Y_n should be uniformly integrable. Since $x^{(i)}$ is bounded, this is obvious.

---

4. Exponential families

[Definition 1] exponential family:

Let $\mathcal{X}$ be a discrete set. Let $\phi: \mathcal{X} \to \mathbb{R}^d$ be a function.

Define a family of distributions $\mathcal{P}$:

$$\mathcal{P} := \{p_\theta : \theta \in \mathbb{R}^d\}, \quad p_\theta(x) := \exp\{\theta \cdot \phi(x) - A(\theta)\}$$

where the log-partition function $A(\theta) := \log \sum_{x\in\mathcal{X}} \exp\{\theta\cdot\phi(x)\}$ ensures the distribution is normalized. (i.e., $\sum p_\theta = 1$)

[Property of exponential family]

1. Gradient and mean:

$$\nabla A(\theta) = E_\theta[\phi(x)] := \sum_{\mathcal{X}} p_\theta(x)\phi(x) \qquad \text{(easy to check)}$$

2. Covariance and Hessian matrix:

$$\nabla^2 A(\theta) = \mathrm{Cov}_\theta[\phi(x)] := E_\theta[(\phi(x)-E_\theta[\phi(x)])(\phi(x)-E_\theta[\phi(x)])^T]$$

Pf: $\nabla(\nabla A(\theta))$

$= \nabla\left(\sum_x p_\theta(x)\phi(x)\right)$

$= \sum_x \nabla p_\theta(x)(\phi(x))^T$

$= \sum_x p_\theta(x)\left(\phi(x) - E_\theta[\phi(x)]\right)\left(\phi(x)\right)^T$

$$= \sum_x P_\theta(x)(\phi(x) - E_\theta[\phi(x)])(\phi(x) - E_\theta[\phi(x)])^T = Cov_\theta[\phi(x)].$$

Since $\sum_x P_\theta(x)(\phi(x) - E_\theta[\phi(x)])(E_\theta[\phi(x)])^T$

$$= \left(\sum_x P_\theta \phi(x) - E_\theta[\phi(x)]\right)(E_\theta[\phi(x)])^T$$

$$= 0$$

Note: (i) since $\nabla^2 A(\theta)$ is a covariance matrix, it is necessarily positive semidefinite, which means that A is convex.

(ii) If $\nabla^2 A(\theta) \succ 0$, then A is strongly convex and ∇A is invertible. In this case, P is said to be minimal.

(iii) If P is minimal, there is a one-to-one mapping:

$$\theta = (\nabla A)^{-1}(\mu), \qquad \mu = \nabla A(\theta)$$

For parameter estimation:

assume $\{x^{(1)}, \cdots, x^{(n)}\} \sim P_{\theta^*}$, i.i.d., the classic way to estimate the distribution is Maximum Likelihood:

$$\hat{P} = P_{\hat{\theta}}, \qquad \hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmax}} \sum_{i=1}^n \log P_\theta(x^{(i)})$$

i.e. $\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmax}} \{\hat{\mu} \cdot \theta - A(\theta)\}$, where $\hat{\mu} := \frac{1}{n} \sum_{i=1}^n \phi(x^{(i)})$

We try to get a close form expression for $\hat{\theta}$ as a function of $\hat{\mu}$.

$\nabla_\theta(\hat{\mu} \cdot \theta - A(\theta)) = \hat{\mu} - \nabla A(\theta)$, since $\hat{\theta}$ is maximal,

$$\Rightarrow \quad \hat{\mu} - \nabla A(\hat{\theta}) = 0$$

$$\Rightarrow \quad \hat{\theta} = (\nabla A)^{-1}(\hat{\mu})$$

Asymptotic analysis:

$$\sqrt{n}(\hat{\mu} - \mu^*) \xrightarrow{d} N(0, Cov_\theta[\phi(x)])$$

where $\mu^* = E[\phi(x)]$.

where $\mu^* = \mathbb{E}_{\theta^*}[\phi(x)]$.

Define $f = (\nabla A)^{-1}$, we have

$$\sqrt{n}(\hat{\theta} - \theta^*) = \sqrt{n}(f(\hat{\mu}) - f(\mu^*))$$

By delta method:

$$\sqrt{n}(f(\hat{\mu}) - f(\mu^*)) \xrightarrow{d} \mathcal{N}(0, \nabla f(\mu^*)^T \mathrm{Cov}_\theta[\phi(x)] \nabla f(\mu^*))$$

Since $\nabla f(\mu) = \nabla^2 A(\mu)^{-1} = \mathrm{Cov}_\theta[\phi(x)]^{-1}$, we have

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathrm{Cov}_\theta[\phi(x)]^{-1})$$

Note: If features vary more, we can estimate $\hat{\theta}$ better.