The moment method and exponential families

John Duchi

Stats 300b – Winter Quarter 2021

Outline

- Moment estimators
- Inverse function theorem
- Exponential family models

Reading: van der Vaart, Chapter 4

Moment method

- function $f : \mathcal{X} \to \mathbb{R}^d$ with $P\|f\|^2 < \infty$, $P_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$,

$$\sqrt{n}(P_n f - Pf) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

  for $\Sigma = \mathrm{Cov}(f)$
- parameter θ of parametric family $\{P_\theta\}_{\theta \in \Theta}$ of interest
- expectation mapping $e : \Theta \to \mathbb{R}^d$ with

$$e(\theta) := \mathbb{E}_\theta[f(X)] = P_\theta f$$

- basic idea: use e^{-1} to estimate θ

Moment method: heuristic

- if e is really smooth, then $(\dot{e^{-1}}) = \frac{\partial}{\partial t} e^{-1}(t)$ exists at $t = P_\theta f$
- delta method gives asymptotics of

$$\sqrt{n}\left(e^{-1}(P_n f) - e^{-1}(Pf)\right)$$

The inverse function theorem

Lemma (cf. van der Vaart Lemmas 4.2–4.3)

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be continuously differentiable in a neighborhood of $\theta \in \mathbb{R}^d$ with invertible Jacobian $F'(\theta) \in \mathbb{R}^{d \times d}$. Then in a neighborhood of $t = F(\theta)$, the derivative

$$(F^{-1})'(t) = \frac{\partial}{\partial t} F^{-1}(t) = (F'(F^{-1}(t)))^{-1}$$

exists and is continuous

The moment method

Theorem

Let $e(\theta) := P_\theta f$ *be one-to-one on an open set* $\Theta \subset \mathbb{R}^d$ *and continuously differentiable at* $\theta_0 \in \Theta$ *with nonsingular* $e'(\theta_0) \in \mathbb{R}^{d \times d}$*. Assume* $P_{\theta_0} \|f\|^2 < \infty$ *and* $X_i \overset{\text{iid}}{\sim} P_{\theta_0}$*. Then* $P_n f \in \operatorname{dom} e^{-1}$ *eventually, and* $\widehat{\theta}_n = e^{-1}(P_n f)$ *satisfies*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \underset{P_{\theta_0}}{\overset{d}{\to}} \mathcal{N}\left(0, e'(\theta_0)^{-1} \mathrm{Cov}_{\theta_0}(f) e'(\theta_0)^{-1}\right)$$

Bernoulli estimation

Example (Bernoullis in $\{\pm 1\}$)

Parameterize by $p_\theta(x) = \frac{e^{\theta x}}{1+e^{\theta x}} = \frac{1}{1+e^{-\theta x}}$. For $e(\theta) = \mathbb{E}_\theta[X]$,

$$\sqrt{n}(e^{-1}(P_n X) - \theta) \stackrel{d}{\to} \mathcal{N}\left(0, \frac{4}{p_\theta(1-p_\theta)}\right)$$

Exponential Family Models

the main example for success of moment methods

Definition

A family $\{P_\theta\}_{\theta \in \Theta}$ is a (regular) *exponential family* with respect to a base measure μ on $\mathcal{X}$ if there exists $T : \mathcal{X} \to \mathbb{R}^d$ and P_θ has density

$$p_\theta(x) = \exp(\theta^\top T(x) - A(\theta)) \text{ w.r.t. } \mu,$$
$$A(\theta) := \log \int \exp(\theta^\top T(x)) d\mu(x)$$

Example

Normal distribution $X \sim \mathcal{N}(\theta, \sigma^2)$ has
$d\mu(x) = \exp(-\frac{x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)),\ A(\theta) = \frac{1}{2\sigma^2}\theta^2,\ T(x) = \frac{1}{\sigma^2}x.$

The log-partition function

$A(\theta) = \int \exp(\theta^\top T(x)) d\mu(x)$ is the *log partition function*

Theorem

$A(\theta)$ is convex in θ, $\mathcal{C}^\infty$, and for $k \in \mathbb{N}$ and $\alpha \in \mathbb{N}^d$ with $\alpha^\top 1 = k$,

$$\begin{aligned} \frac{\partial^k}{\partial \theta_1^{\alpha_1} \cdots \partial \theta_d^{\alpha_d}} \exp(A(\theta)) &= \int T_1(x)^{\alpha_1} \cdots T_d(x)^{\alpha_d} \exp(\theta^\top T(x)) d\mu(x) \\ &= e^{A(\theta)} \mathbb{E}_\theta[T_1(X)^{\alpha_1} \cdots T_d(X)^{\alpha_d}]. \end{aligned}$$

Useful consequencs and moment equalities

- $\nabla A(\theta) = \mathbb{E}_\theta[T]$

- $\nabla^2 A(\theta) = \mathsf{Cov}_\theta(T)$

- if $e(\theta) = \mathbb{E}_\theta[T]$, then $e'(\theta) = \mathsf{Cov}_\theta(T) = \nabla^2 A(\theta) \succeq 0$

Maximum likelihood in exponential families

Corollary

For

$$L_n(\theta) := -P_n \log p_\theta(X),$$

the MLE $\widehat{\theta}_n = \text{argmin}_\theta L_n(\theta) = e^{-1}(P_n T)$

Asymptotics of MLE in exponential familes

Theorem

If the exponential family $\{P_\theta\}$ is full rank (i.e. $\nabla^2 A(\theta) \succ 0$) then the the MLE $\widehat{\theta}_n$

1. *is (eventually) the unique solution to $P_\theta T = P_n T$ in θ*
2. *satisfies*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \underset{P_{\theta_0}}{\overset{d}{\to}} \mathcal{N}\left(0, \nabla^2(A\theta_0)^{-1}\right) \overset{\text{dist}}{=} \mathcal{N}\left(0, I(\theta_0)^{-1}\right).$$

Example: linear regression

- model $p_\theta(y \mid x) \propto \exp(-\frac{1}{2\sigma^2}(y - x^\top\theta)^2)$, i.e. $Y \mid X = x \sim \mathcal{N}(\theta^\top x, \sigma^2)$
- Fisher information matrix becomes

$$I(\theta) = \frac{1}{\sigma^2}\mathbb{E}[xx^\top]$$