Reading: VDV Chapter 4; TPE Chapter 2.5

Outline of lecture 4:

1. Moment method
   (a) Implicit function theorems
   (b) Exponential family models

2. Some thoughts on Fisher information
   (a) Information inequality (Cramer-Rao)
   (b) The real actual information inequality

1 Recap

1.1 Recap of Taylor expansions
For a vector-valued function $f : \mathbb{R}^d \to \mathbb{R}^d$, we have
\[ f(y) = f(x) + Df(x)(y - x) + O(||y - x||). \]
We can also write
\[ f(y) = f(x) + (Df(x) + E(x,y))(y - x), \]
where $E(x,y) = o(1)$.
If $Df(x)$ is $L$-Lipchitz, we have that
\[ E(x,y) \leq \frac{L}{2} ||y - x||. \]

1.2 Recap of MLE
We denote by $\hat{\theta}_n$ the MLE for $\{P_{\theta}\}$, then (here, $\theta \in \Theta \subset \mathbb{R}^d$)
\[ \sqrt{n}(\hat{\theta}_n - \theta) \overset{d}{\to} N(0, I_\theta^{-1}), \]
where $I_\theta$ is the Fisher information matrix.
2 Moment method

Let $X_1, \ldots, X_n$ be a sample of random variable $X$ from a distribution $P_\theta$ that depends on a parameter $\theta$. Suppose $X$ takes values in $\mathcal{X}$, and that $f : \mathcal{X} \to \mathbb{R}^d$ is a vector-valued function such that $P_\theta||f||^2 < \infty$, we denote by

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

the expectation of $f(X)$ under $P_\theta$.

The idea of moment method is to estimate $\theta$ by

$$\hat{\theta} = \mathbb{P}_n f(X),$$

where

$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^n f(X_i).$$

The starting point of moment method is central limit theorem. For function $f$, we have that

$$\sqrt{n}(\mathbb{P}_n f - P_\theta f) \xrightarrow{d} N(0, \Sigma),$$

where

$$\Sigma = \text{Cov}(f).$$

Suppose $e$ is "really nice", we have that

$$\hat{e} = e^{-1}(\mathbb{P}_n f).$$

We denote by

$$e^{-1}(t) = \frac{\partial}{\partial t} (e^{-1}(t)),$$

and delta method gives that

$$\sqrt{n}(e^{-1}(\mathbb{P}_n f - \theta) = \sqrt{n}(e^{-1}(\mathbb{P}_n f) - e^{-1}(P_\theta f)) \xrightarrow{d} e^{-1}(P_\theta f)N(0, \text{Cov}_\theta f)$$

$$= N(0, (e^{-1})^T(P_\theta f)\text{Cov}_\theta f(e^{-1})(P_\theta f)^T).$$

2.1 Inverse function theorem

**Lemma 1** (VDV Lemmas 4.2-4.3). Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a vector-valued function. We assume that $F$ is continuously differentiable in a neighborhood of $\theta \in \mathbb{R}^d$, and that $F'(\theta) \in \mathbb{R}^{d \times d}$ is invertible for $t$ near $F(\theta)$. Then we have that $F^{-1}(t)$ is well-defined and that

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

2.2 Asymptotic normality via inverse function theorem

In this part, we assume that $P_\theta f = 0$.

**Theorem 2.** Let $e(\theta) = P_\theta f$ be one-to-one on an open set $\Theta \subset \mathbb{R}^d$ and continuously differentiable at $\theta_0 \in \Theta$. Assume $e'(\theta_0) \in \mathbb{R}^{d \times d}$ is non-singular. Assume $P_{\theta_0}||f||^2 < \infty$, $X_i \overset{i.i.d.}{\sim} P_{\theta_0}$, then $\hat{\theta}_n = e^{-1}(P_n f)$ exists eventually, and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d,P_{\theta_0}} N(0, e'(\theta_0)^{-1}P_{\theta_0} f f^T(e'(\theta_0)^{-1})^T).$$
Proof  We have that
\[ P_n f \xrightarrow{a.s.} P_{\theta_0} f = e(\theta_0). \]
Eventually, \( \hat{\theta} = e^{-1}(P_n f) \) exists, and in this neighborhood, \( e^{-1} \) is differentiable with
\[ (e^{-1})'(e(\theta_0)) = \left( e'(e^{-1}(e(\theta_0))) \right)^{-1} = e'(\theta_0)^{-1}. \]

3  Exponential family models

Definition 3.1. \( \{P_{\theta}\}_{\theta \in \Theta} \) is a regular exponential family if there is a sufficient statistic \( T : \mathcal{X} \rightarrow \mathbb{R}^d \) such that \( P_{\theta} \) has density
\[ P_{\theta} = \exp(\theta^T T(x) - A(\theta)) \]
with respect to \( \mu \), where \( A(\theta) = \log \int e^{\theta^T T(x)} d\mu(x) \).

Differentiability of \( A(\theta) \)  \( A(\theta) \) is convex in \( \theta \) and \( C^\infty \) in its domain with
\[ \frac{\partial^k e^{A(\theta)}}{\partial \theta_1^{\alpha_1} \cdots \partial \theta_d^{\alpha_d}} = \int T_1(x)^{\alpha_1} \cdots T_d(x)^{\alpha_d} e^{\theta^T T(x)} d\mu(x) \]
for \( \alpha \in \mathbb{N}^d \), \( \sum_{j=1}^d \alpha_j = k \).

Therefore,
\[ \nabla A(\theta) = \nabla \log e^{A(\theta)} \]
\[ = \frac{1}{e^{A(\theta)}} \int T(x)e^{\theta^T T(x)} d\mu(x) \]
\[ = \mathbb{E}_{\theta}[T(x)], \]
\[ \nabla^2 A(\theta) = \int T T^T dP_{\theta} \]
\[ = \left( \int T dP_{\theta} \right) \left( \int T dP_{\theta} \right)^T \]
\[ = \text{Cov}_{\theta}(T). \]

Applying inverse function theorem  We have
\[ e(\theta) = \mathbb{E}_{\theta}[T(x)], \]
\[ e'(\theta) = \text{Cov}_{\theta}[T(x)]. \]

Assuming \( \text{Cov}_{\theta}[T(x)] \succ 0 \), the solution \( \hat{\theta}_n \) to
\[ \frac{1}{n} \sum_{i=1}^n T(X_i) = e(\theta) = \mathbb{E}_{\theta}[T(x)] \]

3
eventually exists, and
\[
\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow{d} N(0, (e'(\theta_0))^{-1}Cov_{\theta_0}(T(x))(e'(\theta_0))^{-1})^T = N(0, Cov_{\theta_0}(T)^{-1}) = N(0, \mathbb{E}_{\theta_0}(i_{\theta_0}^{T})) = N(0, I_{\theta_0}^{-1}).
\]

Now we show MLE estimator equals moment estimator for exponential families. MLE maximizes \(\theta^T P_n T(x) - A(\theta)\). As
\[
\nabla_{\theta} (\theta^T P_n T(x) - A(\theta)) = P_n T(x) - e(\theta),
\]
we have that MLE estimator \(\hat{\theta}\) is determined by
\[
P_n T(x) = e(\hat{\theta}).
\]

4 Fisher information and the biggest con in the history of statistics

Recall the Fisher information \(I_{\theta} = \mathbb{E}_{\theta}[\nabla l_{\theta}(\nabla l_{\theta})^T]\). Given enough smoothness,
\[
I_{\theta} = -\mathbb{E}[\nabla^2 l_{\theta}].
\]

It seems like larger \(I_{\theta}\) will lead to easier estimation.

4.1 Multi-dimensional information inequalities

The idea is to lower bound the variance of different procedures. Consider \(\delta : \mathcal{X} \to \mathbb{R}\) and \(\Psi : \mathcal{X} \to \mathbb{R}^d\). Suppose that \(\mathbb{E}_{\theta}[\Psi] = 0\). We define \(\gamma = [Cov(\Psi, \delta)]_{i=1}^d \in \mathbb{R}^d\), \(C = Cov_{\theta}(\Psi) = \mathbb{E}_{\theta}[\Psi \Psi^T] \in \mathbb{R}^{d \times d}\).

Lemma 3. We have that
\[
Var(\delta) \geq \gamma^T C^{-1} \gamma.
\]

Proof Consider
\[
Cov(\delta, v^T \Psi) = \mathbb{E}[(\delta - \mathbb{E}\delta)(v^T \Psi)] \leq \sqrt{Var(\delta)} \sqrt{Var(v^T \Psi)}.
\]
\[
Cov(\delta, v^T \Psi) = \sum_{j=1}^d v_j Cov(\delta, \Psi_j) = \sum_{j=1}^d v_j \gamma_j = v^T \gamma.
\]
\[
Var(v^T \Psi) = v^T C v.
\]

We have
\[
\frac{(v^T \gamma)^2}{v^T C v} \leq Var(\delta).
\]

Now we choose \(v\) to optimize the lower bound.

Fact If \(A > 0\), then
\[
\sup_{v \neq 0} \frac{(v^T u)^2}{v^T A v} = u^T A^{-1} u.
\]
Proof of fact

\[ v^T u = (A^{-\frac{1}{2}} v)^T (A^{-\frac{1}{2}} u), \]

\[ v^T A v = ||A^{-\frac{1}{2}} v||^2_2. \]

\[ \frac{(v^T u)^2}{v^T A v} = \frac{[(A^{-\frac{1}{2}} v)^T (A^{-\frac{1}{2}} u)]^2}{||A^{-\frac{1}{2}} v||^2_2} \leq ||A^{-\frac{1}{2}} u||^2_2 = u^T A^{-1} u. \]

The equality holds if \( v = A^{-1} u. \)

Choosing \( v = C^{-1} \gamma \), we gain from the fact that

\[ \text{Var}(\delta) \geq \gamma^T C^{-1} \gamma. \]

\[ \square \]

Theorem 4 (Cramer-Rao). Let \( g(\theta) = \mathbb{E}_\theta[\delta] \in \mathbb{R} \) and \( I_\theta = \mathbb{E}_\theta[\nabla l_\theta(\nabla l_\theta)^T] \succ 0 \), then

\[ \text{Var}_\theta(\delta) \geq (g(\theta))^T I_\theta^{-1} g(\theta). \]

Proof. Set \( \Psi(x) = \nabla l_\theta(x) \), we have that \( \mathbb{E}_\theta[\Psi] = 0 \), and that

\[ \mathbb{E}[(\delta - g(\theta)) \Psi] = \mathbb{E}[\delta \Psi] \]

\[ = \mathbb{E}[\delta \nabla l_\theta] \]

\[ = \mathbb{E}[\delta \nabla \frac{p_\theta}{p_\theta}] \]

\[ = \int \delta \nabla \frac{p_\theta}{p_\theta} d\mu(x). \]

Under good regularity conditions, we have that

\[ \mathbb{E}[(\delta - g(\theta)) \Psi] = \nabla \int \delta(x)p_\theta(x)d\mu(x) = \nabla g(\theta). \]

We take

\[ \gamma = \nabla g(\theta), C = I_\theta \]

to get the desired result.

\[ \square \]

Corollary 5 (Cramer-Rao). If \( \hat{\theta} : \mathcal{X} \to \Theta \) is unbiased, then

\[ \mathbb{E}[||\hat{\theta} - \theta||^2] \geq \text{tr}(I_\theta^{-1}) \]

and

\[ \mathbb{E}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)] \succeq I_\theta^{-1}. \]
Proof. Take
\[ g(\theta) = v^T \theta \]
\[ \delta = v^T \hat{\theta}(X). \]
Applying the Cramer-Rao theorem,
\[ \mathbb{E}[(v^T(\hat{\theta} - \theta))^2] \geq v^T I_\theta^{-1} v \]
and
\[ \mathbb{E}[(v^T(\hat{\theta} - \theta))^2] = \mathbb{E}[tr((\hat{\theta} - \theta)(\hat{\theta} - \theta)^T vv^T)] = v^T \text{Cov}(\hat{\theta}) v. \]

Why this is a con?

1. Proof does not give much intuition.
2. There are tons of great biased estimators.

We have that
\[ \mathbb{E}[(\hat{\theta} - \theta)^2] = (\mathbb{E}(\hat{\theta} - \theta))^2 + \text{Var}(\hat{\theta}). \]
For Gaussian mean estimation, let \( X \sim N(\mu, I_n) \), then the James-Stein estimator \( \hat{\mu} = (1 - \lambda ||X||)X \) is biased, but has lower MSE when \( n \geq 3 \).

For ridge regression, \( y = X\beta + \epsilon \), then the ridge regression estimator is \( \hat{\beta}_\lambda = (X^T X + \lambda I)^{-1} X^T y \), and it has lower MSE than \( \hat{\beta}_{OLS} = \hat{\beta}_0 \) if \( X^T X \) is ill-conditioned.

4.2 The real theorem: Le Cam and Hajek’s local asymptotic minimax theorem
Fix \( \theta_0 \) and let \( \Pi_{n,c} \) be uniform distribution over \( \{\theta : ||\theta - \theta_0|| \leq \frac{c}{\sqrt{n}}\} \). Then for any symmetric, bounded, bowl-shaped \( L \),
\[ \lim_{C \to +\infty} \lim_{n \to +\infty} \lim_{\hat{\theta}_n} \int \mathbb{E}_\theta[L(\sqrt{n}(\hat{\theta}_n - \theta))] \Pi_{n,c}(\theta) d\theta \geq \mathbb{E}[L(Z)], \]
where \( Z \sim N(0, I_{\theta_0}^{-1}) \).

Here, \( \mathbb{E}[L(Z)] \) estimates \( Z \sim N(0, I_{\theta_0}^{-1}) \) by 0. If we let \( L(t) = t^2 \), \( \mathbb{E}[L(Z)] = 2 tr(I_{\theta_0}^{-1}). \)