STAT 460/560 Class 13: Asymptotic Normality of MLEs (redux)

Ben Bloem-Reddy

Reading: Chapter 5.5, [van98].

1. Differentiable in quadratic mean

A statistical model $\{P_\theta : \theta \in \Theta\}$ is **differentiable in quadratic mean** (DQM) if there exists a measurable vector-valued function $\dot{\ell}_{\theta_0}$ such that

$$\int \left[\sqrt{p_\theta(x)} - \sqrt{p_{\theta_0}(x)} - \frac{1}{2}(\theta - \theta_0)^\top \dot{\ell}_{\theta_0}(x) \sqrt{p_{\theta_0}(x)} \right]^2 \mu(dx) = o(\|\theta - \theta_0\|^2) \, . \tag{13.1}$$

Let's unpack this a little bit and establish some identities that will be useful below.

First, notice the implicit assumption that each element of the model has a density (PDF) with respect to the dominating measure μ. When X takes values in $\mathbb{R}^d$, this is typically Lebesgue measure. But it doesn't have to be. For our purposes, just remember that $\int f(x) p_\theta(x) \mu(dx) = \int f(x) P_\theta(dx) = E_\theta[f((X)]$.

Now, what does it mean to be DQM? Recall that if $\Theta \subset \mathbb{R}^k$, a function $\phi \colon \Theta \to \mathbb{R}^m$ is differentiable at θ if there is a matrix $\phi'_\theta \colon \Theta \to \mathbb{R}^m$ such that

$$\phi(\theta + h) - \phi(\theta) = \phi'_\theta(h) + o(\|h\|) \, .$$

Comparing that to (13.1), we see that the DQM condition can be interpreted as follows.

- First notice that $\nabla_\theta \sqrt{p_\theta(x)} = \frac{1}{2} \sqrt{p_\theta(x)} \nabla_\theta \log p_\theta$, so $\dot{\ell}_\theta = \nabla_\theta \log p_\theta$, which is the score function as usual.
- Suppose that as a function of θ, $\sqrt{p_\theta(x)}$ (for fixed x) is differentiable at θ_0. Here, the function $\phi(\theta) = \sqrt{p_\theta(x)}$, which takes values in $\mathbb{R}$ (so $m = 1$) and ϕ'_{θ_0} is just the vector-valued function $\dot{\ell}_{\theta_0}$ (with $h = \theta - \theta_0$).
- Now ask whether we have differentiability of $\sqrt{p_\theta(x)}$ at each possible value of x. Then the LHS integral splits over two sets, A_{diff}), where differentiability holds, and its complement

  $$\begin{aligned} &o(\|\theta - \theta_0\|^2)\mu(A_{\text{diff}}) + \int_{A^c_{\text{diff}}} \left[\sqrt{p_\theta(x)} - \sqrt{p_{\theta_0}(x)} - \frac{1}{2}(\theta - \theta_0)^\top \dot{\ell}_{\theta_0}(x) \sqrt{p_{\theta_0}(x)} \right]^2 \mu(dx) \\ &\geq o(\|\theta - \theta_0\|^2)\mu(A_{\text{diff}}) + C\mu(A^c_{\text{diff}}) \, , \end{aligned}$$

  where $C > 0$ is some constant that results from non-differentiability on the set A^c_{diff}.

So, we see that if a model is DQM then $\sqrt{p_\theta(x)}$ is differentiable (as a function of θ) at θ_0 for μ-almost all points $x \in \mathcal{X}$. Because μ is a common dominating measure for the model, this means that if $X \sim P_\theta$ then the random function $\sqrt{p_\theta(X)}$ will be differentiable with probability 1.

So what does this have to do with proving asymptotic normality of MLEs? Ultimately, we would like to apply Theorem 5.23 in [van98], so we need to use DQM to show that using the log-likelihood of a model that is DQM satisfies the the conditions of the theorem. To do so, we will show that $\log p_\theta$ is **differentiable in P-probability** at θ_0, which means that for every sequence $\theta_n \to \theta$, for every $\epsilon > 0$,

$$\lim_{n \to \infty} P\left(\left| \log p_{\theta_n}(X) - \log p_{\theta_0}(X) - (\theta_n - \theta_0)^\top \dot{\ell}_{\theta_0}(X) \right| > \epsilon \right) \to 0 \, . \tag{13.2}$$

We will also use the DQM property to show that $\log p_{\theta_0}$ admits a second-order Taylor expansion, which was the other main requirement for Theorem 5.23.

We're almost ready for the theorem and proof. But let's derive a couple of identities that we'll refer to in the proof. First, for a sequence of vectors $h_n \to h \in \Theta$, define $\theta_n := \theta_0 + h_n/\sqrt{n}$, and

$$W_n(X) := 2\left(\sqrt{\frac{p_{\theta_n}(X)}{p_{\theta_0}(X)}} - 1\right) .$$

Then

$$\log p_{\theta_n}(X) - \log p_{\theta_0}(X) = 2\log\left(1 + \frac{1}{2}W_n(X)\right) . \tag{13.3}$$

Moreover, DQM can be written as

$$\int \left(\sqrt{n}\frac{1}{2}W_n(x) - \frac{1}{2}h_n^\top \dot{\ell}_{\theta_0}(x)\right)^2 p_{\theta_0}(x)\mu(dx) = o(\|h_n^2\|) = o(1) . \tag{13.4}$$

Re-writing as an expectation, we see that

$$E_{\theta_0}\left[\left(\sqrt{n}W_n(X) - h_n^\top \dot{\ell}_{\theta_0}(X)\right)^2\right] \to 0 ,$$

and therefore $\sqrt{n}W_n(X) \xrightarrow{\text{qm}} h^\top \dot{\ell}_{\theta_0}(X)$, which implies that $\sqrt{n}W_n(X) \xrightarrow{P_{\theta_0}} h^\top \dot{\ell}_{\theta_0}(X)$.

Finally going back to the DQM criterion, we see that the sequence $\sqrt{n}(\sqrt{p_{\theta_n}(X)} - \sqrt{p_{\theta_0}(X)})$ converges to $\frac{1}{2}h^\top \dot{\ell}_{\theta_0}$ in $L_2(\mu)$.

Now suppose that $h \to 0$. Then

$$\int \left(\sqrt{p_{\theta_n}(x)} - \sqrt{p_{\theta_0}(x)}\right)^2 \mu(dx) \to 0 ,$$

implying that $\sqrt{p_{\theta_n}(X)} \to \sqrt{p_{\theta_0}(X)}$ in $L_2(\mu)$.

Finally,

$$\begin{aligned}
P_{\theta_0} h^\top \dot{\ell}_{\theta_0} &= \int \frac{1}{2} h^\top \dot{\ell}_{\theta_0}(x)\sqrt{p_{\theta_0}(x)}2\sqrt{p_{\theta_0}(x)}\mu(dx) \\
&= \lim_n \int \sqrt{n}(\sqrt{p_{\theta_n}(x)} - \sqrt{p_{\theta_0}(x)})(\sqrt{p_{\theta_n}(x)} + \sqrt{p_{\theta_0}(x)})\mu(dx) \\
&= \lim_n \int \sqrt{n} p_{\theta_n}(x)\mu(dx) - \int \sqrt{n} p_{\theta_0}(x)\mu(dx) = 0 .
\end{aligned}$$

We'll use all of these in the proof below.

Theorem 13.1. *Suppose that the model $\{P_\theta : \theta \in \Theta\}$ is DQM at an inner point θ_0 of $\Theta \subset \mathbb{R}^k$. Suppose that there exists a measurable function $\dot{\ell}$ with $P_{\theta_0}\dot{\ell}^2 < \infty$ such that in a neighborhood of θ_0,*

$$|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \le \dot{\ell}(x)\|\theta_1 - \theta_2\| .$$

If the Fisher information matrix $I_{\theta_0} = P(\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}^\top)$ is nonsingular and $\hat{\theta}_n \xrightarrow{p} \theta_0$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = I_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) + o_{P_{\theta_0}}(1) , \tag{13.5}$$

from which asymptotic normality (with mean zero and covariance matrix $I_{\theta_0}^{-1}$) follows.

Proof. Above, we showed that $\sqrt{n}W_n(X) \xrightarrow{P_{\theta_0}} h^\top \dot{\ell}_{\theta_0}(X)$. By the continuous mapping theorem, we also have that $nW_n(X)^2 \xrightarrow{P_{\theta_0}} h^\top \dot{\ell}_{\theta_0}(X)\dot{\ell}_{\theta_0}(X)^\top h$. Because $\sqrt{n}W_n(X) \xrightarrow{\text{qm}} h^\top \dot{\ell}_{\theta_0}(X)$, we also have that $E_{\theta_0}[nW_n(X)^2] \to h^\top I_{\theta_0} h$.

By (13.3),

$$\begin{aligned}\sqrt{n}(\log p_{\theta_n}(X) - \log p_{\theta_0}(X)) &= 2\sqrt{n}\log\left(1 + \frac{1}{2}W_n(X)\right) \\ &= 2\sqrt{n}\left(\frac{1}{2}W_n(X) - \frac{1}{8}W_n(X)^2 + W_n(X)^2 R(W_n(X))\right) ,\end{aligned}$$

where R is a function satisfying $R(w) \to 0$ as $w \to 0$ (the second equality and the function R follow from the Taylor expansion of $\log(1+w) = w - \frac{1}{2}w^2 + w^2 R(w)$ around $w = 0$). Therefore,

$$\sqrt{n}(\log p_{\theta_n}(X) - \log p_{\theta_0}(X)) = h^\top \dot{\ell}_{\theta_0}(X) + o_{P_{\theta_0}}(1) .$$

The higher-order terms converge to zero in probability because: $\sqrt{n}W_n(X) = nW_n(X)/\sqrt{n} = ((h^\top \dot{\ell}_{\theta_0}(X))^2 + o_P(1))/\sqrt{n} = (O_P(1) + o_P(1))/\sqrt{n} \xrightarrow{\text{p}} 0$. The second term is $O_P(1)o_P(1) = o_P(1)$. This shows that the function $\theta \mapsto \log p_\theta$ is differentiable in P_{θ_0}-probability at θ_0. (Which is one of the necessary conditions for Theorem 5.23.)

The above convergence in probability can be strengthened to convergence in quadratic mean.[1]

We still need to show that $\log p_{\theta_0}$ admits a second-order Taylor expansion. To that end,

$$\begin{aligned}P_{\theta_0}(nW_n) &= 2n\left(\int \sqrt{\frac{p_{\theta_n}(x)}{p_{\theta_0}(x)}} p_{\theta_0}(x)\mu(dx) - 1\right) - 0 \\ &= 2n\left(\int \sqrt{p_{\theta_n}(x)}\sqrt{p_{\theta_0}(x)}\mu(dx) - 1\right) \\ &= -n\left(\int (\sqrt{p_{\theta_n}(x)} - \sqrt{p_{\theta_0}(x)})^2 \mu(dx)\right) \to -\frac{1}{4}h^\top P_{\theta_0}(\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}^\top)h = -\frac{1}{4}h^\top I_{\theta_0} h .\end{aligned}$$

Now, using the Taylor series approximation of $\log(1+w)$ again,

$$\begin{aligned}nP_{\theta_0}(\log p_{\theta_n} - \log p_{\theta_0}) &= 2nE_{\theta_0}[\log(1 + \frac{1}{2}W_n(X))] \\ &= E_{\theta_0}[nW_n(X)] - \frac{1}{4}E_{\theta_0}[nW_n(X)^2] + E_{\theta_0}[nW_n(X)^2 R(W_n(X))] \\ &\to -\frac{1}{2}h^\top I_{\theta_0} h .\end{aligned}$$

This establishes the required second-order Taylor expansion. Hence, Theorem 5.23 applies. □

References

[van98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[1] Roughly, the argument is to use the Lipschitz condition and the dominated convergence theorem to show that the sequence $n(\log p_{\theta_n}(X) - \log p_{\theta_0}(X))^2$ is uniformly integrable, which with the convergence in probability, implies that the convergence also holds in quadratic mean.