Lecture 7: MLE in generalized linear models (GLM) and quasi-MLE

MLE in exponential families

Suppose that X has a distribution from a natural exponential family so that the likelihood function is

$$\ell(\eta) = \exp\{\eta^\tau T(x) - \zeta(\eta)\} h(x),$$

where $\eta \in \Xi$ is a vector of unknown parameters.
The likelihood equation is then

$$\frac{\partial \log \ell(\eta)}{\partial \eta} = T(x) - \frac{\partial \zeta(\eta)}{\partial \eta} = 0,$$

which has a unique solution $T(x) = \partial \zeta(\eta)/\partial \eta$, assuming that $T(x)$ is in the range of $\partial \zeta(\eta)/\partial \eta$.
Note that

$$\frac{\partial^2 \log \ell(\eta)}{\partial \eta \partial \eta^\tau} = -\frac{\partial^2 \zeta(\eta)}{\partial \eta \partial \eta^\tau} = -\mathrm{Var}(T)$$

(see the proof of Proposition 3.2).

Since $\text{Var}(T)$ is positive definite, $-\log \ell(\eta)$ is convex in η and $T(x)$ is the unique MLE of the parameter $\mu(\eta) = \partial\zeta(\eta)/\partial\eta$.
Also, the function $\mu(\eta)$ is one-to-one so that μ^{-1} exists.
By Definition 4.3, the MLE of η is $\widehat{\eta} = \mu^{-1}(T(x))$.

If the distribution of X is in a general exponential family and the likelihood function is

$$\ell(\theta) = \exp\{[\eta(\theta)]^\tau T(x) - \xi(\theta)\}h(x),$$

then the MLE of θ is $\widehat{\theta} = \eta^{-1}(\widehat{\eta})$, if η^{-1} exists and $\widehat{\eta}$ is in the range of $\eta(\theta)$.
Of course, $\widehat{\theta}$ is also the solution of the likelihood equation

$$\frac{\partial \log \ell(\theta)}{\partial\theta} = \frac{\partial\eta(\theta)}{\partial\theta} T(x) - \frac{\partial\xi(\theta)}{\partial\theta} = 0.$$

Suppose that $X_1, ..., X_n$ are i.i.d. with a distribution in a natural exponential family, i.e., the p.d.f. of X_i is

$$f_\eta(x_i) = \exp\{\eta^\tau T(x_i) - \zeta(\eta)\}h(x_i).$$

From Proposition 3.2 and $\partial^2 \log f_\eta(x_i)/\partial\eta\partial\eta^\tau = -\partial^2\zeta(\eta)/\partial\eta\partial\eta^\tau$, all conditions in Theorem 4.16 are satisfied.

If $\widehat{\theta}_n = n^{-1}\sum_{i=1}^n T(X_i) \in \Theta$, the range of $\theta = g(\eta) = \partial\zeta(\eta)/\partial\eta$, then $\widehat{\theta}_n$ is a unique RLE of θ, which is also a unique MLE of θ since $\partial^2\zeta(\eta)/\partial\eta\partial\eta^\tau = \mathrm{Var}(T(X_i))$ is positive definite.

Also, $\eta = g^{-1}(\theta)$ exists and a unique RLE (MLE) of η is $\widehat{\eta}_n = g^{-1}(\widehat{\theta}_n)$.

However, $\widehat{\theta}_n$ may not be in Θ and the previous argument fails (e.g., Example 4.29).

What Theorem 4.17 tells us in this case is that as $n \to \infty$, $P(\widehat{\theta}_n \in \Theta) \to 1$ and, therefore, $\widehat{\theta}_n$ (or $\widehat{\eta}_n$) is the unique asymptotically efficient RLE (MLE) of θ (or η) in the limiting sense.

In an example like this we may directly show that $P(\widehat{\theta}_n \in \Theta) \to 1$, using the fact that $\widehat{\theta}_n \to_{a.s.} E[T(X_1)] = g(\eta)$ (the SLLN).

The results for exponential families lead to an estimation method in a class of models that have very wide applications.

Generalized linear models (GLM)

The GLM is a generalization of the normal linear model discussed in §3.3.1-§3.3.2.

The GLM is useful since it covers situations where the relationship between $E(X_i)$ and Z_i is nonlinear and/or X_i's are discrete.

The structure of a GLM

The sample $X = (X_1, ..., X_n)$ has independent X_i's and X_i has the p.d.f.

$$\exp\left\{\frac{\eta_i x_i - \zeta(\eta_i)}{\phi_i}\right\} h(x_i, \phi_i), \qquad i = 1, ..., n,$$

w.r.t. a σ-finite measure ν, where η_i and ϕ_i are unknown, $\phi_i > 0$,

$$\eta_i \in \Xi = \left\{\eta : \ 0 < \int h(x, \phi) e^{\eta x/\phi} d\nu(x) < \infty\right\} \subset \mathscr{R}$$

for all i, ζ and h are known functions, and $\zeta''(\eta) > 0$ is assumed for all $\eta \in \Xi^\circ$, the interior of Ξ.

Note that the p.d.f. belongs to an exponential family if ϕ_i is known.

As a consequence,

$$E(X_i) = \zeta'(\eta_i) \quad \text{and} \quad \mathrm{Var}(X_i) = \phi_i \zeta''(\eta_i), \qquad i = 1, ..., n.$$

Define $\mu(\eta) = \zeta'(\eta)$.

It is assumed that η_i is related to Z_i, the ith value of a p-vector of covariates, through

$$g(\mu(\eta_i)) = \beta^\tau Z_i, \qquad i = 1, ..., n,$$

where β is a p-vector of unknown parameters and g, called a *link function*, is a known one-to-one, third-order continuously differentiable function on $\{\mu(\eta) : \eta \in \Xi^\circ\}$.

If $\mu = g^{-1}$, then $\eta_i = \beta^\tau Z_i$ and g is called the *canonical* or *natural* link function.

If g is not canonical, we assume that $\frac{d}{d\eta}(g \circ \mu)(\eta) \neq 0$ for all η.
In a GLM, the parameter of interest is β.

We assume that the range of β is

$$B = \{\beta : (g \circ \mu)^{-1}(\beta^\tau z) \in \Xi^\circ \text{ for all } z \in \mathscr{Z}\}$$

where $\mathscr{Z}$ is the range of Z_i's.

ϕ_i's are called *dispersion* parameters and are considered to be nuisance parameters.

MLE in GLM

An MLE of β in a GLM is considered under assumption

$$\phi_i = \phi/t_i, \qquad i = 1,...,n,$$

with an unknown $\phi > 0$ and known positive t_i's.
Let $\theta = (\beta, \phi)$ and $\psi = (g \circ \mu)^{-1}$.

$$\log \ell(\theta) = \sum_{i=1}^{n} \left[\log h(x_i, \phi/t_i) + \frac{\psi(\beta^\tau Z_i)x_i - \zeta(\psi(\beta^\tau Z_i))}{\phi/t_i} \right]$$

$$\frac{\partial \log \ell(\theta)}{\partial \beta} = \frac{1}{\phi} \sum_{i=1}^{n} \{[x_i - \mu(\psi(\beta^\tau Z_i))]\psi'(\beta^\tau Z_i)t_i Z_i\} = 0$$

$$\frac{\partial \log \ell(\theta)}{\partial \phi} = \sum_{i=1}^{n} \left\{ \frac{\partial \log h(x_i, \phi/t_i)}{\partial \phi} - \frac{t_i[\psi(\beta^\tau Z_i)x_i - \zeta(\psi(\beta^\tau Z_i))]}{\phi^2} \right\} = 0.$$

From the first likelihood equation, an MLE of β, if it exists, can be obtained without estimating ϕ.
The second likelihood equation, however, is usually difficult to solve.
Some other estimators of ϕ are suggested by various researchers; see, for example, McCullagh and Nelder (1989).

Suppose that there is a solution $\widehat{\beta} \in B$ to the likelihood equation.

$$\mathrm{Var}\left(\frac{\partial \log \ell(\theta)}{\partial \beta}\right) = M_n(\beta)/\phi, \quad \frac{\partial^2 \log \ell(\theta)}{\partial \beta \partial \beta^\tau} = [R_n(\beta) - M_n(\beta)]/\phi.$$

where

$$M_n(\beta) = \sum_{i=1}^n [\psi'(\beta^\tau Z_i)]^2 \zeta''(\psi(\beta^\tau Z_i)) t_i Z_i Z_i^\tau$$

$$R_n(\beta) = \sum_{i=1}^n [x_i - \mu(\psi(\beta^\tau Z_i))] \psi''(\beta^\tau Z_i) t_i Z_i Z_i^\tau.$$

Consider first the simple case of canonical g, $\psi'' \equiv 0$ and $R_n \equiv 0$.
If $M_n(\beta)$ is positive definite for all β, then $-\log \ell(\theta)$ is strictly convex in β for any fixed ϕ and, therefore, $\widehat{\beta}$ is the unique MLE of β.
For noncanonical g, $R_n(\beta) \neq 0$ and $\widehat{\beta}$ is not necessarily an MLE.
If $R_n(\beta)$ is dominated by $M_n(\beta)$, i.e.,

$$[M_n(\beta)]^{-1/2} R_n(\beta) [M_n(\beta)]^{-1/2} \to 0$$

in some sense, then $-\log \ell(\theta)$ is convex and $\widehat{\beta}$ is an MLE for large n.
In a GLM, an MLE $\widehat{\beta}$ usually does not have an analytic form and a numerical method such as the Newton-Raphson has to be applied.

Example 4.36

Consider the GLM with $\zeta(\eta) = \eta^2/2$, $\eta \in \mathscr{R}$.
If g is the canonical link, then the model is the same as a linear model with independent ε_i's distributed as $N(0, \phi_i)$.
If $\phi_i \equiv \phi$, then the likelihood equation is exactly the same as the normal equation in §3.3.1.
If Z is of full rank, then $M_n(\beta) = Z^\tau Z$ is positive definite.
Thus, the LSE $\widehat{\beta}$ in a normal linear model is the unique MLE of β.
Suppose now that g is noncanonical but $\phi_i \equiv \phi$.
Then the model reduces to the one with independent X_i's and

$$X_i = N\left(g^{-1}(\beta^\tau Z_i), \phi\right), \qquad i = 1, ..., n.$$

This type of model is called a *nonlinear regression model* (with normal errors) and an MLE of β under this model is also called a nonlinear LSE, since maximizing the log-likelihood is equivalent to minimizing the sum of squares $\sum_{i=1}^n [X_i - g^{-1}(\beta^\tau Z_i)]^2$.
Under certain conditions the matrix $R_n(\beta)$ is dominated by $M_n(\beta)$ and an MLE of β exists.

Example 4.37 (The Poisson model)

Consider the GLM with $\zeta(\eta) = e^{\eta}$, $\eta \in \mathscr{R}$, $\phi_i = \phi / t_i$.
If $\phi_i = 1$, then X_i has the Poisson distribution with mean e^{η_i}.
Under the canonical link $g(t) = \log t$,

$$M_n(\beta) = \sum_{i=1}^{n} e^{\beta^\tau Z_i} t_i Z_i Z_i^\tau,$$

which is positive definite if $\inf_i e^{\beta^\tau Z_i} > 0$ and the matrix $(\sqrt{t_1} Z_1, ..., \sqrt{t_n} Z_n)$ is of full rank.
There is one noncanonical link that deserves attention.
Suppose that we choose a link function so that $[\psi'(t)]^2 \zeta''(\psi(t)) \equiv 1$.
Then $M_n(\beta) \equiv \sum_{i=1}^{n} t_i Z_i Z_i^\tau$ does not depend on β.
In §4.5.2 it is shown that the asymptotic variance of the MLE $\widehat{\beta}$ is $\phi [M_n(\beta)]^{-1}$.
The fact that $M_n(\beta)$ does not depend on β makes the estimation of the asymptotic variance (and, thus, statistical inference) easy.
Under the Poisson model, $\zeta''(t) = e^t$ and, therefore, we need to solve the differential equation $[\psi'(t)]^2 e^{\psi(t)} = 1$.
A solution is $\psi(t) = 2\log(t/2)$ and the link $g(\mu) = 2\sqrt{\mu}$.

Theorem 4.18

Consider the GLM with $\phi_i = \phi/t_i$ and t_i's in a fixed interval (t_0, t_∞), $0 < t_0 \le t_\infty < \infty$.
Assume that the range of the unknown parameter β is an open subset of $\mathscr{R}^p$; at the true value of β, $0 < \inf_i \varphi(\beta^\tau Z_i) \le \sup_i \varphi(\beta^\tau Z_i) < \infty$, where $\varphi(t) = [\psi'(t)]^2 \zeta''(\psi(t))$; as $n \to \infty$, $\max_{i \le n} Z_i^\tau (Z^\tau Z)^{-1} Z_i \to 0$ and $\lambda_-[Z^\tau Z] \to \infty$, where Z is the $n \times p$ matrix whose ith row is the vector Z_i and $\lambda_-[A]$ is the smallest eigenvalue of A.

(i) There is a unique sequence of estimators $\{\widehat{\beta}_n\}$ such that

$$P\left(s_n(\widehat{\beta}_n) = 0\right) \to 1 \qquad \text{and} \qquad \widehat{\beta}_n \to_p \beta,$$

where $s_n(\beta) = \partial \log \ell(\beta, \phi)/\partial \beta$ is the score function.

(ii) Let $I_n(\beta) = \mathrm{Var}(s_n(\beta))$. Then

$$[I_n(\beta)]^{1/2}(\widehat{\beta}_n - \beta) \to_d N_p(0, I_p).$$

(iii) If ϕ is known or the p.d.f. indexed by $\theta = (\beta, \phi)$ satisfies the conditions for f_θ in Theorem 4.16, then $\widehat{\beta}_n$ is asymptotically efficient.

Key issues in the proof of Theorem 4.18

The proof of asymptotic existence and consistency is similar to that of Theorem 4.17.

For the asymptotic normality of $\widehat{\beta}_n$, we still use Taylor's expansion and, similar to the proof of Theorem 4.17, can establish that

$$[I_n(\beta)]^{1/2}(\widehat{\beta}_n - \beta) = [I_n(\beta)]^{-1/2} s_n(\beta) + o_p(1),$$

where $I_n(\beta) = M_n(\beta)/\phi$.

Using the CLT (e.g., Corollary 1.3) and Theorem 1.9(iii), we can show (exercise) that

$$[I_n(\beta)]^{-1/2} s_n(\beta) \to_d N_p(0, I_p).$$

These two results and Slutsky's theorem imply that

$$[I_n(\beta)]^{1/2}(\widehat{\beta}_n - \beta) \to_d N(0, I_p)$$

Since $I_n(\beta)$ is the Fisher information about β, this result implies that $\widehat{\beta}_n$ is asymptotically efficient when ϕ is known.

Key issues in the proof of Theorem 4.18

When ϕ is unknown, however, we cannot directly conclude from the previous result whether $\widehat{\beta}_n$ is asymptotically efficient.

A complete argument for the asymptotic efficiency of $\widehat{\beta}_n$ is as follows. Note that

$$\frac{\partial}{\partial \phi}\left[\frac{\partial \log \ell(\theta)}{\partial \beta}\right] = -\frac{s_n(\beta)}{\phi}.$$

Since $E[s_n(\beta)] = 0$, the Fisher information about $\theta = (\beta, \phi)$ is

$$I_n(\beta, \phi) = -E\left[\frac{\partial^2 \log \ell(\theta)}{\partial \theta \partial \theta^\tau}\right] = \begin{pmatrix} I_n(\beta) & 0 \\ 0 & \tilde{I}_n(\phi) \end{pmatrix},$$

where $\tilde{I}_n(\phi)$ is the Fisher information about ϕ.

Then the asymptotic efficiency of $\widehat{\beta}_n$ follows from

$$[I_n(\beta, \phi)]^{-1} = \begin{pmatrix} [I_n(\beta)]^{-1} & 0 \\ 0 & [\tilde{I}_n(\phi)]^{-1} \end{pmatrix}$$

Quasi-MLE

If assumption ϕ_i is arbitrary, or the distribution assumption on X_i does not hold (e.g., X_i is longitudinal), but

$$E(X_i) = \zeta'(\eta_i) \quad \text{and} \quad \mathrm{Var}(X_i) = \phi_i \zeta''(\eta_i), \qquad i = 1, ..., n.$$

and

$$g(\mu(\eta_i)) = \beta^\tau Z_i, \qquad i = 1, ..., n,$$

still hold, and we estimate β by solving equation

$$G_n(\beta) = \sum_{i=1}^{n} \left\{ [x_i - \mu(\psi(\beta^\tau Z_i))] \psi'(\beta^\tau Z_i) t_i Z_i \right\} = 0$$

then the resulting estimator is called a quasi-MLE.

This method is also called the method of generalized estimating equations (GEE).

They are efficient if the GEE is a likelihood equation, and is robust if it is not.

Quasi-MLE or GEE has some good asymptotic properties.

Discussion of asymptotic properties of quasi-MLE

The asymptotic existence and consistency of quasi-MLE can be shown using a similar argument to the proof of Theorem 4.17.

To show the asymptotic normality, using the Taylor expansion we obtain that

$$-G_n(\beta) = \nabla G_n(\beta)(\widehat{\beta}_n - \beta) + o_p(n^{-1/2})$$

Then

$$-\sqrt{n}[\nabla G_n(\beta)]^{-1} G_n(\beta) = \sqrt{n}(\widehat{\beta}_n - \beta) + o_p(1)$$

By the SLLN and CLT,

$$n^{-1}\nabla G_n(\beta) \to_{a.s.} \Gamma \qquad n^{-1/2} G_n(\beta) \to_d N(0, \Sigma)$$

where $\Sigma = \mathrm{Var}(G_n(\beta))$ and Γ is a positive definite matrix.
Hence,

$$\begin{aligned}\sqrt{n}(\widehat{\beta}_n - \beta) &= -\sqrt{n}[\nabla G_n(\beta)]^{-1} G_n(\beta) + o_p(1) \\ &\to_d N(0, \Gamma^{-1}\Sigma\Gamma^{-1})\end{aligned}$$

If $\widehat{\beta}_n$ is an MLE, then $\Gamma = \Sigma =$ Fisher information and $\Gamma^{-1}\Sigma\Gamma^{-1} = \Sigma^{-1}$.