Statistical Inference: MLE Theory

Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch

October 25, 2025

Motivation

- provides a general paradigm for inference on parametric models, with many generalisations and variants;
- uses only minimal sufficient statistics;
- is a central concept in both frequentist and Bayesian statistics;
- has a simple, general and widely-applicable 'large-sample' theory; but
- is not a panacea!

- Let $Y, Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g$, and define the **Kullback–Leibler divergence** from the **data-generating model** g to a **candidate density** f,

$$\mathrm{KL}(g, f) = \mathrm{E}_g\{\log g(Y) - \log f(Y)\} = \mathrm{E}_g\left[-\log\left\{\frac{f(Y)}{g(Y)}\right\}\right] \geq 0,$$

  using the fact that $-\log x$ is convex and we can apply Jensen's inequality. The inequality is strict unless $f \equiv g$.

- In a parametric setting f belongs to a parametric family $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, so minimising $\mathrm{KL}(g, f)$ over f is equivalent to maximising $\mathrm{E}_g \log f(Y; \theta)$, which is estimated by

$$\overline{\ell}(\theta) = n^{-1} \sum_{j=1}^{n} \log f(Y_j; \theta) \xrightarrow{P} \mathrm{E}_g \log f(Y; \theta), \quad n \to \infty.$$

- $\theta_g = \mathrm{argmax}_\theta \mathrm{E}_g \log f(Y; \theta)$ gives the optimal large-sample fit of f_θ to g.
- In an ideal case $g \in \mathcal{F}$, so $g = f_{\theta_g}$, but the theory does not require this (yet).
- The natural estimator of θ_g is the **maximum likelihood estimator**

$$\widehat{\theta} = \mathrm{argmax}_\theta \, \overline{\ell}(\theta),$$

  but we need conditions on $\overline{\ell}$ to ensure that $\widehat{\theta} \xrightarrow{P} \theta_g$ or (better) $\widehat{\theta} \xrightarrow{\text{a.s.}} \theta_g$ as $n \to \infty$.

Example (MLE for Gaussian distribution)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The likelihood is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(Y_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum_{i=1}^{n}(Y_i-\mu)^2}{2\sigma^2}\right\}.$$

giving loglikelihood

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2.$$

All partial second derivatives exist and are

$$\begin{aligned}
\frac{\partial}{\partial\mu}\ell(\mu, \sigma^2) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu) \\
\frac{\partial}{\partial\sigma^2}\ell(\mu, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - \mu)^2.
\end{aligned}$$

Example (MLE for Gaussian distribution, continued)

Solving $\nabla_{(\mu,\sigma^2)}\ell(\mu,\sigma^2) = 0$ for (μ,σ^2) gives a system of equations in two unknowns, with unique root

$$\left(\overline{Y}, n^{-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right).$$

Call this $(\widehat{\mu},\widehat{\sigma}^2)$, and let's verify it's a maximum. Note that

$$\frac{\partial^2}{\partial\mu^2}\ell(\mu,\sigma^2) = -\frac{n}{\sigma^2}, \quad \frac{\partial^2}{\partial(\sigma^2)^2}\ell(\mu,\sigma^2) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(Y_i-\mu)^2$$

$$\frac{\partial^2}{\partial\mu\partial\sigma^2}\ell(\mu,\sigma^2) = \frac{\partial^2}{\partial\sigma^2\partial\mu}\ell(\mu,\sigma^2) = -\frac{\sum_{i=1}^{n}(Y_i-\mu)}{\sigma^4} = \frac{n\mu - n\overline{Y}}{\sigma^4}.$$

Calculating these derivatives at $(\widehat{\mu},\widehat{\sigma}^2)$, we get

$$\left.\frac{\partial^2}{\partial\mu^2}\ell(\mu,\sigma^2)\right|_{(\mu,\sigma^2)=(\widehat{\mu},\widehat{\sigma}^2)} = -\frac{n}{\widehat{\sigma}^2}, \quad \left.\frac{\partial^2}{\partial(\sigma^2)^2}\ell(\mu,\sigma^2)\right|_{(\mu,\sigma^2)=(\widehat{\mu},\widehat{\sigma}^2)} = -\frac{n}{2\widehat{\sigma}^4}$$

Example (MLE for Gaussian distribution, continued)

$$\frac{\partial^2}{\partial\mu\partial\sigma^2}\ell(\mu,\sigma^2)\Big|_{(\mu,\sigma^2)=(\widehat{\mu},\widehat{\sigma}^2)} = \frac{\partial^2}{\partial\sigma^2\partial\mu}\ell(\mu,\sigma^2)\Big|_{(\mu,\sigma^2)=(\widehat{\mu},\widehat{\sigma}^2)} = \frac{n\widehat{\mu}-n\widehat{\mu}}{\widehat{\sigma}^4} = 0.$$

Thus the matrix

$$\left[-\nabla^2_{(\mu,\sigma^2)}\ell(\mu,\sigma^2)\Big|_{(\mu,\sigma^2)=(\widehat{\mu},\widehat{\sigma}^2)}\right]$$

is diagonal. If both of its diagonal elements are positive, then it will be positive definite. This is indeed the case since $\widehat{\sigma}^2 > 0$ nd so the unique MLE of (μ,σ^2) is given by

$$(\widehat{\mu},\widehat{\sigma}^2) = \left(\overline{Y}\ ,\ \frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right).$$

□

Note that from Gaussian sampling results we get that σ^2 is biased.

Example (MLE for Poisson Distribution)

Let $Y_1, ..., Y_n \overset{iid}{\sim} \text{Poisson}(\lambda)$. Then

$$L(\lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{Y_i}}{Y_i!} e^{-\lambda} \right\} \implies \log L(\lambda) = -n\lambda + \log \lambda \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \log(Y_i!)$$

Setting $\nabla_\lambda \log L(\lambda) = -n + \lambda^{-1} \sum Y_i = 0$ we obtain $\widehat{\lambda} = \overline{Y}$ since $\nabla^2_\lambda \log L(\lambda) = -\lambda^{-2} \sum Y_i < 0$.

Example (MLE for Uniform Distribution – a non-differentiable case)

Let $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$. The likelihood is

$$L(\theta) = \theta^{-n} \prod_{i=1}^{n} \mathbf{1}\{0 \leq Y_i \leq \theta\} = \theta^{-n} \mathbf{1}\{\theta \geq Y_{(n)}\}.$$

Hence if $\theta < Y_{(n)}$ the likelihood is zero. In the domain $[Y_{(n)}, \infty)$, the likelihood is a decreasing function of θ. Hence $\widehat{\theta} = Y_{(n)}$.

Example (Equivariance of the MLE)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$, and suppose we're interested in estimating $\mathbb{P}[Y_1 \leq y]$, for a given $y \in \mathbb{R}$. Note that

$$\mathbb{P}[Y_1 \leq y] = \mathbb{P}[Y_1 - \mu \leq y - \mu] = \Phi(y - \mu),$$

where Φ is the standard normal CDF. The mapping $\mu \mapsto \Phi(y - \mu)$ is bijective, since Φ is strictly monotone. So by equivariance, the MLE of $\mathbb{P}[Y_1 \leq y]$ is $\Phi(y - \widehat{\mu})$, where $\widehat{\mu}$ is the MLE of μ (which by our previous example is $\widehat{\mu} = \overline{Y}$).

Example (Equivariance and usual vs natural parameterisation)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} f$, with

$$f(y) = \exp\{\phi T(y) - \gamma(\phi) + S(y)\}, \qquad y \in \mathcal{Y}$$

where $\phi \in \Phi \subseteq \mathbb{R}$ is the natural parameter. Suppose we can write $\phi = \eta(\theta)$, where $\theta \in \Theta$ is the usual parameter and $\eta : \Theta \to \Phi$ is a differentiable bijection (so that $\gamma(\phi) = \gamma(\eta(\theta)) = d(\theta)$, for $d = \gamma \circ \eta$). In this notation, the density/frequency takes the form

$$\exp\{\phi T(y) - \gamma(\phi) + S(y)\} = \exp\{\eta(\theta) T(y) - d(\theta) + S(y)\}.$$

Equivariance now implies that if $\widehat{\theta}$ is the MLE of θ, then $\eta(\widehat{\theta})$ is the MLE of $\phi = \eta(\theta)$. The converse is also true: if $\widehat{\phi}$ is the MLE of ϕ, then $\eta^{-1}(\widehat{\phi})$ is the MLE of $\theta = \eta^{-1}(\phi)$. □

Examples show that likelihood generally gives sensible estimators – still:

- Beyond intuition, is there a canonical mathematical reason for it?
- What rigorous guarantees can we offer?
  - ↪ Can we get consistency?
  - ↪ Can we approach reasonable MSE performance?

To answer these questions, we go back to entropy and Kullback-Leibler divergence.

Consider the random function

$$\Psi_n(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n}[\log f(Y_i;\boldsymbol{u}) - \log f(Y_i;\boldsymbol{\theta})]$$

which is maximized at $\widehat{\boldsymbol{\theta}}_n$. By the law of large numbers, for each $\boldsymbol{u} \in \Theta$,

$$\Psi_n(\boldsymbol{u}) \xrightarrow{p} \Psi(\boldsymbol{u}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\log\left(\frac{f(Y_i;\boldsymbol{u})}{f(Y_i;\boldsymbol{\theta})}\right)\right] = -KL\big(f(Y_i;\boldsymbol{u})\big\|f(Y_i;\boldsymbol{\theta})\big)$$

- The latter is minimised at $\boldsymbol{\theta}$ and so $\Psi(u)$ is maximized at $\boldsymbol{\theta}$.
- Moreover, unless $f(x;\boldsymbol{u}) = f(x;\boldsymbol{\theta})$ for all $x \in \operatorname{supp} f$, we have $\Psi(\boldsymbol{u}) < 0$
- It follows that Ψ is uniquely maximised at $\boldsymbol{\theta}$

MLE can be regarded as a minimiser of an approximate (empirically constructed) KL-divergence from the truth!

Does $\left\{\Psi_n(\boldsymbol{u}) \xrightarrow{p} \Psi(\boldsymbol{u}) \ \forall \ u \text{ with } \Psi \text{ maximized uniquely at } \boldsymbol{\theta}\right\}$ imply $\left\{\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}\right\}$?

- Unfortunately, the answer is in general no, without additional information.
- If $\theta \in \mathbb{R}$, can prove consistency if f is regular enough & MLE exists uniquely.
- If $\theta \in \mathbb{R}^p$, we need more information on the form of the likelihood function
  - $\hookrightarrow$ For instance concavity and existence will usually give us consistency. We will show consistency in exponential families using this approach.
  - $\hookrightarrow$ More general situations require stronger forms of convergence of $\Psi_n(u) \to \Psi(u)$ plus additional regularity conditions.

When we can deduce consistency, though, we get some very nice properties for the (asymptotic) sampling distribution of the MLE...

Example (Consistency of MLE in $\theta \in \mathbb{R}$)

Let $Y_1, ..., Y_n \overset{iid}{\sim} f(y;\theta)$ where f is C^1 with respect to θ. Assume that $\forall\ n$, there exists a unique MLE $\widehat{\theta}_n$. We will show that $\widehat{\theta}_n \xrightarrow{p} \theta$.

Define

$$\Xi_n(u) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial u}\log\left(\frac{f(Y_i;u)}{f(Y_i;\theta)}\right)\right] \quad \text{and} \quad \Xi(u) = \mathbb{E}\left[\frac{\partial}{\partial u}\log\left(\frac{f(Y_i;u)}{f(Y_i;\theta)}\right)\right],$$

so that

- $\Xi_n(\widehat{\theta}_n) = 0$ uniquely, by assumption.
- $\Xi(\theta) = 0$ uniquely, assuming regularity allowing interchange of $\mathbb{E}$ and $\frac{\partial}{\partial u}$.

Since f is C^1, we have the inequality

$$\mathbb{P}[\Xi_n(\theta-\varepsilon) < 0 \ \&\ \Xi_n(\theta+\varepsilon) > 0] \leq \mathbb{P}[\theta-\varepsilon < \widehat{\theta}_n < \theta+\varepsilon]$$

because the event on the left hand side implies that on the right hand side. Finally, the law of large numbers implies that $\Xi_n(u) \xrightarrow{p} \Xi(u)$ for any u, so that the left hand side converges to 1, yielding consistency.

Example (Consistency of MLE in $\mathbb{R}^k$ for exponential families)

Consider $Y_1, ..., Y_n \overset{iid}{\sim} f(y; \phi)$ from a k-parameter exponential family

$$f(y) = \exp\left\{\sum_{j=1}^{k} \phi_j T_j(y) - \gamma(\phi_1, ..., \phi_k) + S(y)\right\}, \boldsymbol{\phi} = (\phi_1, ..., \phi_k)^\top \in \Phi \text{ open.}$$

The likelihood and loglikelihood (up to constants w.r.t. $\boldsymbol{\phi}$) are given by

$$L(\boldsymbol{\phi}) = \exp\left\{\boldsymbol{\phi}^\top \boldsymbol{\tau} - n\gamma(\boldsymbol{\phi})\right\} \quad \& \quad \ell(\boldsymbol{\phi}) = \boldsymbol{\phi}^\top \boldsymbol{\tau} - n\gamma(\boldsymbol{\phi})$$

where

$$\boldsymbol{\tau} = (\tau_1, ..., \tau_k)^\top, \qquad \tau_j(y_1, \ldots, y_n) = \textstyle\sum_{i=1}^{n} T_j(y_i).$$

If it exists, the MLE $\widehat{\boldsymbol{\phi}}_n$ must thus satisfy

$$\nabla_{\boldsymbol{\phi}} \ell(\widehat{\boldsymbol{\phi}}_n) = 0 \implies \nabla_{\boldsymbol{\phi}} \gamma(\widehat{\boldsymbol{\phi}}_n) = n^{-1}\boldsymbol{\tau}.$$

Furthermore, existence of the MLE guarantees uniqueness by strict concavity:

$$-\nabla^2_{\boldsymbol{\phi}} \ell(\boldsymbol{\phi}) = n\nabla^2_{\boldsymbol{\phi}} \gamma(\boldsymbol{\phi}) = \operatorname{cov}\{\boldsymbol{\tau}\} \succ 0,$$

Example (Consistency of MLE in $\mathbb{R}^k$ for exponential families, ctd)

Now notice that by the law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n} T_j(Y_i) \xrightarrow{p} \mathbb{E}[T_j] = \frac{\partial}{\partial \phi_j}\gamma(\boldsymbol{\phi}), \qquad j = 1, ..., k.$$

It follows that

$$\nabla_{\boldsymbol{\phi}}\gamma(\widehat{\boldsymbol{\phi}}_n) = n^{-1}\boldsymbol{\tau} \xrightarrow{p} \nabla_{\boldsymbol{\phi}}\gamma(\boldsymbol{\phi}).$$

Now if $\nabla_{\boldsymbol{\phi}}\gamma : \mathbb{R}^k \to \mathbb{R}^k$ were continuously invertible, with inverse map h, then the continuous mapping theorem would give us:

$$\nabla_{\boldsymbol{\phi}}\gamma(\widehat{\boldsymbol{\phi}}_n) \xrightarrow{p} \nabla_{\boldsymbol{\phi}}\gamma(\boldsymbol{\phi}) \implies h\big(\nabla_{\boldsymbol{\phi}}\gamma(\widehat{\boldsymbol{\phi}}_n)\big) \xrightarrow{p} h\big(\nabla_{\boldsymbol{\phi}}\gamma(\boldsymbol{\phi})\big) \implies \widehat{\boldsymbol{\phi}}_n \xrightarrow{p} \boldsymbol{\phi}.$$

In fact, the inverse function theorem tells us that the infinitely differentiable function $\nabla_{\boldsymbol{\phi}}\gamma : \mathbb{R}^k \to \mathbb{R}^k$ must admit a continuously differentiable inverse map h locally.

In summary: provided it exists, the MLE of the natural parameter in a k-parameter natural exponential family with open parameter space Φ is consistent.

Assuming we can get consistency, we can focus on understanding the sampling distribution of the MLE.

For simplicity, assume $X_1, ..., X_n$ are iid with density/frequency $f(x;\theta)$, $\theta \in \mathbb{R}$.

Introduce the notation:

- $\ell(x_i;\theta) = \log f(x_i;\theta)$
- $\ell'(x_i;\theta)$, $\ell''(x_i;\theta)$ and $\ell'''(x_i;\theta)$ are partial derivatives w.r.t θ.

Regularity Conditions $(*)$

(A1) Θ is an open subset of $\mathbb{R}$.

(A2) The support of f, $\text{supp}(f)$, is independent of θ.

(A3) f is thrice continuously differentiable w.r.t. θ for all $x \in \text{supp}(f)$.

(A4) $\mathbb{E}_\theta[\ell'(X_i;\theta)] = 0 \ \forall\theta$ and $\text{var}_\theta[\ell'(X_i;\theta)] = \mathcal{I}_1(\theta) \in (0,\infty) \ \forall\theta$.

(A5) $-\mathbb{E}_\theta[\ell''(X_i;\theta)] = \mathcal{J}_1(\theta) \in (0,\infty) \ \forall\theta$.

(A6) $\exists\ M(x) > 0$ and $\delta > 0$ such that $\mathbb{E}_{\theta_0}[M(X_i)] < \infty$ and

$$|\theta - \theta_0| < \delta \implies |\ell'''(x;\theta)| \leq M(x)$$

Let's demistify these conditions...

- If Θ is open, then for θ the true parameter, it always makes sense for an estimator $\widehat{\theta}$ to have a symmetric distribution around θ (e.g. Gaussian).
- Under condition (A2) we have $\frac{d}{d\theta}\int_{\text{supp } f} f(x;\theta)dx = 0$ for all $\theta \in \Theta$ so that, if we can interchange integration and differentiation,

$$0 = \int \frac{d}{d\theta} f(x;\theta)dx = \int \ell'(x;\theta)f(x;\theta)dx = \mathbb{E}_\theta[\ell'(X_i;\theta)]$$

  so that in the presence of (A2), (A4) is essentially a condition that enables differentiation under the integral and asks that the r.v. ℓ' have a finite second moment for all θ.
- Similarly, (A5) requires that ℓ'' have a first moment for all θ.
- Conditions (A2) and (A6) are smoothness conditions that will allow us to 'linearize' the problem, while the other conditions will allow us to 'control' the random linearization.
- Furthermore, if we can differentiate twice under the integral sign

$$0 = \int \frac{d}{d\theta}[\ell'(x;\theta)f(x;\theta)]dx = \int \ell''(x;\theta)f(x;\theta)dx + \int (\ell'(x;\theta))^2 f(x;\theta)dx$$

  so that $\imath(\theta) = \jmath(\theta)$.

Theorem (Asymptotic Distribution of the MLE)

Let $X_1, ..., X_n$ be iid random variables with density (frequency) $f(x;\theta)$ and satisfying the stated regularity conditions. If the MLE $\widehat{\theta}_n$ exists uniquely and is consistent, we have

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\imath_1(\theta)}{\jmath_1^2(\theta)}\right).$$

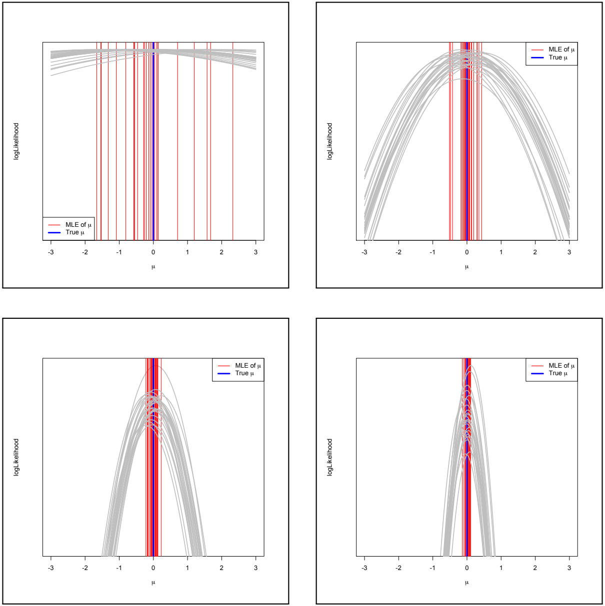
When $\imath_1(\theta) = \jmath_1(\theta)$, we have of course $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\imath_1(\theta)}\right)$.

- Note that this can be interpreted as

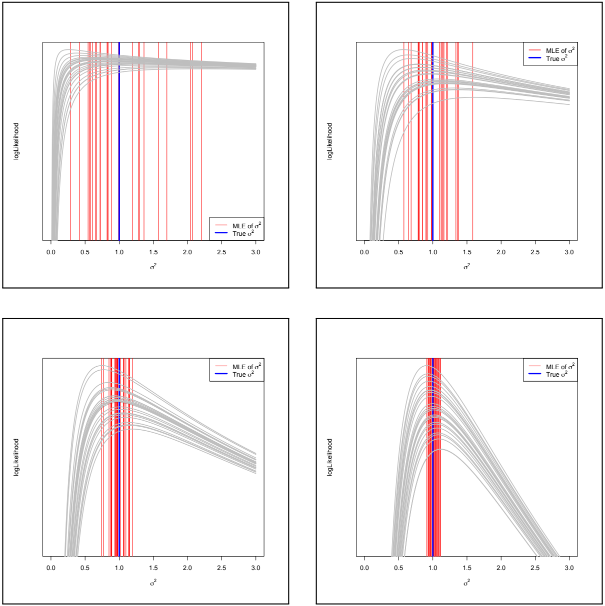
$$\widehat{\theta}_n \overset{d}{\approx} N\left(\theta\,,\, \frac{1}{n\imath_1(\theta)}\right) \equiv N\left(\theta\,,\, \frac{1}{\imath_n(\theta)}\right).$$

- In order words: the MLE is approximately normally distributed, approximately unbiased, and approximately achieving the Cramér-Rao lower bound!

Why $\mathcal{I}_n(\theta)$? (... curvature)



Why $\mathcal{I}_n(\theta)$? (... curvature)



Proof.

Under conditions (A1)-(A3), if $\widehat{\theta}_n$ maximizes the likelihood, we have

$$\sum_{i=1}^{n} \ell'(X_i; \widehat{\theta}_n) = 0.$$

Expanding this equation in a Taylor series, we get

$$\begin{aligned} 0 = \sum_{i=1}^{n} \ell'(X_i; \widehat{\theta}_n) \quad &= \quad \sum_{i=1}^{n} \ell'(X_i; \theta) + \\ &\quad + (\widehat{\theta}_n - \theta) \sum_{i=1}^{n} \ell''(X_i; \theta) \\ &\quad + \tfrac{1}{2} (\widehat{\theta}_n - \theta)^2 \sum_{i=1}^{n} \ell'''(X_i; \theta_n^*) \end{aligned}$$

with θ_n^* lying between θ and $\widehat{\theta}_n$.

Dividing accross by $\sqrt{n}$ yields

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell'(X_i;\theta) + \sqrt{n}(\widehat{\theta}_n - \theta) \frac{1}{n} \sum_{i=1}^{n} \ell''(X_i;\theta) \\ &\quad + \frac{1}{2}\sqrt{n}(\widehat{\theta}_n - \theta)^2 \frac{1}{n} \sum_{i=1}^{n} \ell'''(X_i;\theta_n^*) \end{aligned}$$

which suggests that $\sqrt{n}(\widehat{\theta}_n - \theta)$ equals

$$\frac{-n^{-1/2} \sum_{i=1}^{n} \ell'(X_i;\theta)}{n^{-1} \sum_{i=1}^{n} \ell''(X_i;\theta) + (\widehat{\theta}_n - \theta)(2n)^{-1} \sum_{i=1}^{n} \ell'''(X_i;\theta_n^*)}.$$

Now, from the central limit theorem and condition (A4), it follows that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell'(X_i;\theta) \xrightarrow{d} \mathcal{N}(0, \imath_1(\theta)).$$

Next, the weak law of large numbers along with condition (A5) implies

$$\frac{1}{n}\sum_{i=1}^{n}\ell''(X_i;\theta) \xrightarrow{p} -\mathcal{J}(\theta).$$

By Slutsky's lemma, the theorem will follow if we show that $R_n \xrightarrow{p} 0$. This is established in the next lemma, which we appeal to, completing the proof. □

Lemma

In the same context as in the previous theorem,

$$R_n = (\widehat{\theta}_n - \theta)\frac{1}{2n}\sum_{i=1}^{n}\ell'''(X_i;\theta_n^*) \xrightarrow{p} 0$$

for any random variable θ_n^* on the segment joining $\widehat{\theta}_n$ and θ.

Proof. $(*)$

We have that for any $\epsilon > 0$

$$\mathbb{P}[|R_n| > \epsilon] = \underbrace{\mathbb{P}[|R_n| > \epsilon, |\widehat{\theta}_n - \theta| > \delta]}_{\leq \mathbb{P}[|\widehat{\theta}_n - \theta| > \delta] \xrightarrow{p} 0} + \mathbb{P}[|R_n| > \epsilon, |\widehat{\theta}_n - \theta| \leq \delta]$$

If $|\widehat{\theta}_n - \theta| < \delta$, (A6) implies $|R_n| \leq \frac{\delta}{2n} \sum_{i=1}^{n} M(X_i) = \overline{M}_n$.
so we may write

$$\mathbb{P}[|R_n| > \epsilon, |\widehat{\theta}_n - \theta| \leq \delta] \leq \mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta \overline{M}_n]$$

and for $\xi > 0$, the last term can be bounded by

$$\begin{aligned} &\mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta \overline{M}_n, \overline{M}_n \leq M + \xi] + \\ &\quad + \mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta \overline{M}_n, \overline{M}_n > M + \xi] \end{aligned}$$

which in turn is bounded by

$$\begin{aligned} \leq \quad & \mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta(M + \xi)] + \mathbb{P}[\overline{M}_n > M + \xi] \\ \leq \quad & \mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta(M + \xi)] + \mathbb{P}[|\overline{M}_n - M| > \xi] \end{aligned}$$

But the law of large numbers implies that

$$\overline{M}_n = \frac{1}{n} \sum_{i=1}^{n} M(X_i) \xrightarrow{p} \mathbb{E}[M(X_1)] < \infty,$$

It follows that

$$\mathbb{P}[|\overline{M}_n - M| > \xi] \to 0.$$

Since we can always choose δ to be as small as we wish, we can make the term

$$\mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta(M + \xi)]$$

equal to zero. In summary, we have established that $R_n \xrightarrow{p} 0$

□

- The true model is supposed to lie in the candidate family, i.e., $g \in \mathcal{F}$, so $\theta_g \in \Theta$.
- Mathematically speaking the assumption that $g \in \mathcal{F}$ is always false, but
  - the asymptotic results are supposed to provide guidelines on what to expect when fitting models — checking the regularity conditions in practice would require knowledge of g, in which case there's no need for inference!
  - this is largely irrelevant if model-checking suggests that $f_{\widehat{\theta}_g}$ is 'close enough' to g.
- Crucially, the interest parameter θ should have a stable interpretation for candidates likely to be close to g (i.e., within $n^{-1/2}$) , so $\mathcal{F}$ is 'robustly specified' — if the model is not quite right, then the interpretation of the crucial parameters will be unchanged.

- We usually assume classical asymptotics and replace the variance $\frac{\imath_1(\theta)}{\jmath_1^2(\theta)}$ by the inverse of the **observed information matrix**

$$\widehat{\jmath} = -\nabla^2 \ell(\widehat{\theta}),$$

which

  - can be computed numerically without (possibly awkward) expectations,
  - will (helpfully!) misbehave if the maximisation is questionable,
  - has been found to give generally good results in applications,
  - has the heuristic justification that $(\widehat{\theta}, \widehat{\jmath})$ are approximately sufficient for θ_g, as

$$\ell(\theta_g) \doteq \ell(\widehat{\theta}) - \tfrac{1}{2}(\widehat{\theta} - \theta_g)^{\mathrm{T}} \widehat{\jmath} (\widehat{\theta} - \theta_g).$$

- Standard errors for $\widehat{\theta}$ are the square roots of the diagonal elements of $\widehat{\jmath}^{-1}$.

- Classical asymptotics support inference for scalar θ based on any of the (approximate) pivots

$$T = t(\theta_g) = \widehat{\jmath}^{1/2}(\widehat{\theta} - \theta_g) \dot{\sim} \mathcal{N}(0,1), \qquad \textbf{Wald statistic},$$
$$S = s(\theta_g) = \widehat{\jmath}^{-1/2}\nabla\ell(\theta_g) \dot{\sim} \mathcal{N}(0,1), \qquad \textbf{score statistic},$$
$$W = w(\theta_g) = 2\{\ell(\widehat{\theta}) - \ell(\theta_g)\} \dot{\sim} \chi^2_1, \qquad \textbf{likelihood ratio statistic}.$$

- If $\widehat{\theta}^{\mathrm{o}}$ and $\widehat{\jmath}(\widehat{\theta}^{\mathrm{o}})$ have been obtained for observed data y^{o}, then the approximation

$$\Pr_g\{T(\theta_g) \leq t^{\mathrm{o}}(\theta_g)\} \doteq \Phi\{t^{\mathrm{o}}(\theta_g)\}$$

leads to $(1-\alpha)$ **Wald confidence interval** $\widehat{\theta}^{\mathrm{o}} \pm \jmath(\widehat{\theta}^{\mathrm{o}})^{-1/2} z_{1-\alpha/2}$ based on T, while that based on W is

$$\{\theta : W^{\mathrm{o}}(\theta) \leq \chi^2_1(1-\alpha)\} = \{\theta : \ell^{\mathrm{o}}(\theta) \geq \ell^{\mathrm{o}}(\widehat{\theta}^{\mathrm{o}}) - \tfrac{1}{2}\chi^2_1(1-\alpha)\},$$

where z_p and $\chi^2_\nu(p)$ are respectively the p quantiles of the $N(0,1)$ and χ^2_ν distributions.

Comparative comments

- Confidence intervals based on T are symmetric, but those based on W take the shape of ℓ into account and are parametrisation-invariant;
- in small samples the distributional approximations for W are better than that for T, and that for W can be improved by **Bartlett correction**, using $W_{\mathrm{B}} = W/(1 + b/n)$;
- confidence sets based on W may not be connected (and if so T is unreliable);
- the main use of S is for testing in situations where maximisation of ℓ is awkward.

- The regularity conditions for MLE asymptotics apply in many settings met in practice, but not universally. The most common failures arise when
  - some of the parameters are discrete (e.g., change point problems),
  - the model is not identifiable (distinct θ values give the same model),
  - θ_g is on the boundary of the parameter space (e.g., testing for a zero variance),
  - $d = \dim(\theta)$ grows (too fast) with n, or
  - the support of $f(y;\theta)$ depends on θ (so the Bartlett identities fail).
- Even when the conditions are satisfied there can be datasets for which maximum likelihood estimation fails, e.g.,
  - there is no unique maximum to the likelihood, or
  - the maximum is on the edge of the parameter space,

  and then penalisation (equivalent to using a prior) is often used.

Example

If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} U(0, \theta)$, show that the limit distribution of $n(\theta - \widehat{\theta})/\theta$ when $n \to \infty$ is exp(1).

First, we check the Bartlett identities. In this case $1 = \int f(y;\theta)\,\mathrm{d}y = \int_0^\theta \theta^{-1}\,\mathrm{d}y$, and differentiation with respect to θ gives

$$0 = 1/\theta + \int_0^\theta (-\theta^{-2})\,\mathrm{d}y,$$

so the first Bartlett identity is not satisfied (because the support depends on θ, and $f(\theta;\theta) \neq 0$. The likelihood can still be constructed as

$$L(\theta) = \prod_{j=1}^n f_Y(y_j;\theta) = \prod_{j=1}^n \{\theta^{-1} I(0 < y_j < \theta)\} = \theta^{-n} I(\max y_j < \theta), \quad \theta > 0,$$

and therefore $\widehat{\theta} = M = \max Y_j$, whose distribution is

$$\Pr(M \leq x) = (x/\theta)^n, \quad 0 < x < \theta.$$

Now

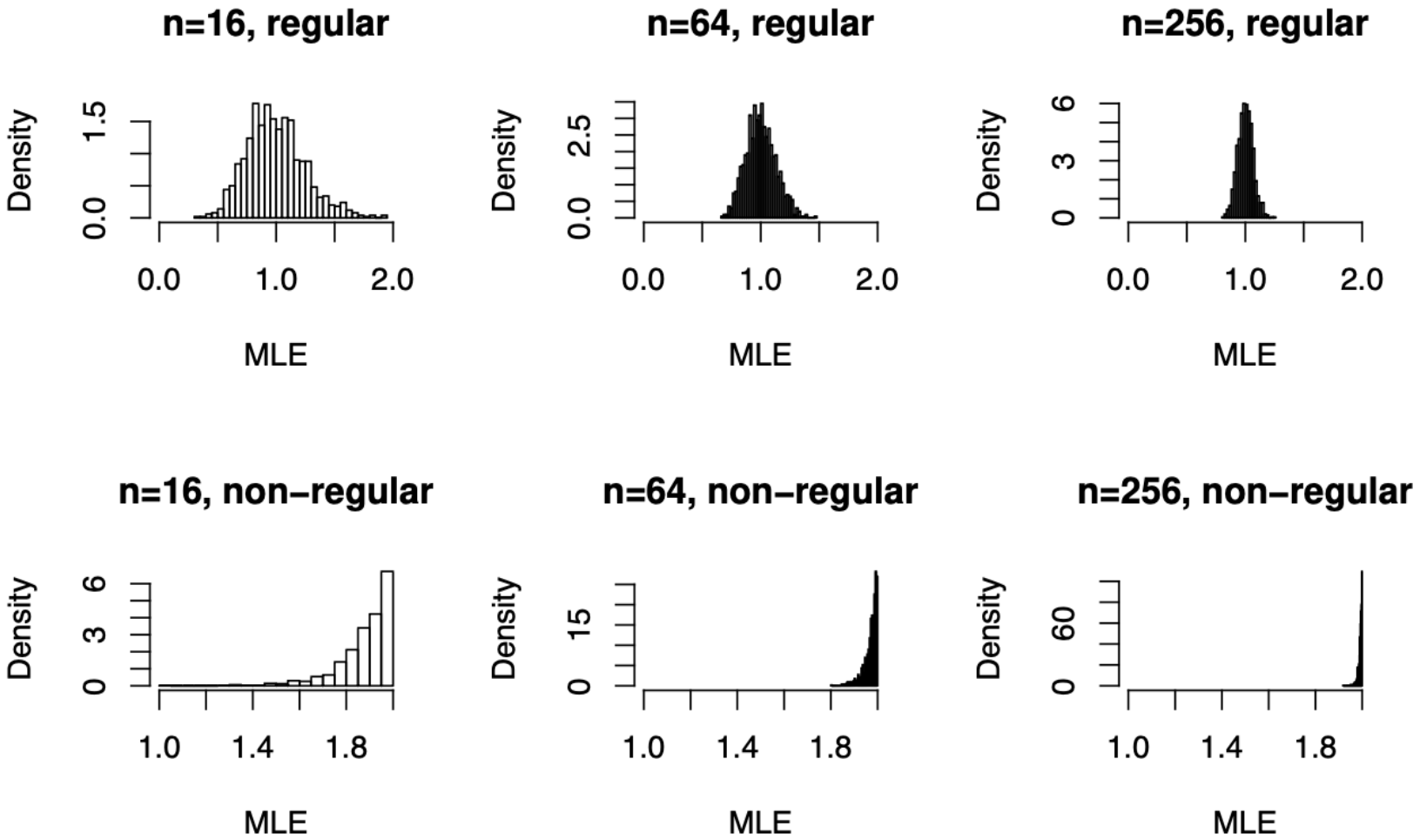
$$\Pr\left\{n(\theta - \widehat{\theta})/\theta \leq x\right\} = \Pr(\widehat{\theta} \geq \theta - x\theta/n) = 1 - \{(\theta - x\theta/n)/\theta\}^n \to 1 - \exp(-x),$$

as required.

Note: the scaling needed to get a limiting distribution is much faster here than in the regular case (we have to multiply by n to get a non-degenerate limit); the limiting distribution is not normal.

Uniform example

Comparison of the distributions of $\widehat{\theta}$ in a regular case (panels above, with standard deviation $\propto n^{-1/2}$) and in a nonregular case , panels below, with standard deviation $\propto n^{-1}$). In other nonregular cases it might happen that the distribution is nasty (unlike here) and/or that the convergence is slower than in regular cases.



If θ divides into a $p \times 1$ **interest parameter** ψ and a $q \times 1$ **nuisance parameter** λ, then

$$\widehat{\theta} = \begin{pmatrix} \widehat{\psi} \\ \widehat{\lambda} \end{pmatrix} \dot{\sim} \mathcal{N}_{p+q} \left\{ \begin{pmatrix} \psi_g \\ \lambda_g \end{pmatrix}, \begin{pmatrix} \widehat{\mathcal{J}}_{\psi\psi} & \widehat{\mathcal{J}}_{\psi\lambda} \\ \widehat{\mathcal{J}}_{\lambda\psi} & \widehat{\mathcal{J}}_{\lambda\lambda} \end{pmatrix}^{-1} \right\},$$

where for brevity we now write $\widehat{\lambda}_\psi = \max_\lambda \ell(\psi, \lambda)$, $\tilde{\theta} = \widehat{\theta}_\psi = (\psi, \widehat{\lambda}_\psi)$,

$$\ell_\psi = \left. \frac{\partial \ell(\theta)}{\partial \psi} \right|_{\theta=\theta_g}, \quad \widehat{\mathcal{J}}_{\psi\psi} = -\widehat{\ell}_{\psi\psi} = - \left. \frac{\partial^2 \ell(\theta)}{\partial \psi \partial \psi^{\mathrm{T}}} \right|_{\theta=\widehat{\theta}}, \quad \tilde{\ell}_{\psi\psi} = \left. \frac{\partial^2 \ell(\theta)}{\partial \psi \partial \psi^{\mathrm{T}}} \right|_{\theta=\tilde{\theta}}, \quad \text{etc.}$$

- Under classical asymptotics and setting $\widehat{\mathcal{J}}^{\psi\psi} = (\widehat{\mathcal{J}}_{\psi\psi} - \widehat{\mathcal{J}}_{\psi\lambda}\widehat{\mathcal{J}}_{\lambda\lambda}^{-1}\widehat{\mathcal{J}}_{\lambda\psi})^{-1}$ we have

$$\widehat{\psi} \dot{\sim} \mathcal{N}_p\left(\psi_g, \widehat{\mathcal{J}}^{\psi\psi}\right) \qquad \textbf{maximum likelihood estimator},$$

$$s(\psi_g) = \tilde{\ell}_\psi^{\mathrm{T}} \widehat{\mathcal{J}}^{\psi\psi} \tilde{\ell}_\psi \dot{\sim} \chi^2_p \qquad \textbf{score statistic},$$

$$w_{\mathrm{p}}(\psi_g) = 2\left\{\ell_{\mathrm{p}}(\widehat{\psi}) - \ell_{\mathrm{p}}(\psi_g)\right\} \dot{\sim} \chi^2_p \qquad \textbf{(generalized) likelihood ratio statistic},$$

  where we defined w_{p} using the **profile log likelihood** $\ell_{\mathrm{p}}(\psi) = \ell(\psi, \widehat{\lambda}_\psi) = \max_\lambda \ell(\psi, \lambda)$.

- Properties:
  - inferences using $w(\psi_g)$ are invariant to interest-respecting reparametrisation, so are preferable but more computationally burdensome;
  - $s(\psi_g)$ is mainly used for tests, since only λ must be estimated (as $\psi = \psi_g$ is known).
- A $(1-\alpha)$ confidence set based on $w_{\mathrm{p}}(\psi_g)$ (or equivalently on $\ell_{\mathrm{p}}(\psi)$) is

$$\left\{\psi : w_{\mathrm{p}}(\psi) \leq \chi^2_p(1-\alpha)\right\} = \left\{\psi : \ell(\psi, \widehat{\lambda}_\psi) \geq \ell(\widehat{\psi}, \widehat{\lambda}) - \tfrac{1}{2}\chi^2_p(1-\alpha)\right\}.$$

- The fact that $\mathrm{KL}(g, f)$ is minimised when $f = g$ suggests comparing competing models $\mathcal{F}_1, \ldots, \mathcal{F}_M$ by their maximised log likelihoods $\log f_m(y; \widehat{\theta}_m) = \widehat{\ell}_m$.
- But $\widehat{\ell}_m$ should be penalized, because
  - $\widehat{\ell}_m \geq \log f_m(y; \theta_m)$ even if $\mathcal{F}_m$ is the true model class, and
  - enlarging θ_m can increase $\widehat{\ell}_m$ even if further parameters are unnecessary.
- Akaike proposed minimising $2\mathrm{E}_g\mathrm{E}_g^+\left[-\log\{f(Y^+; \widehat{\theta})/g(Y^+)\}\right]$, where $Y^+, Y \overset{\text{iid}}{\sim} g$ are independent datasets. The idea is that if $\widehat{\theta} = \widehat{\theta}(Y)$ is estimated separately from Y^+, there will be a penalty due to 'missing θ_g' which will grow with $\dim(\theta)$.
- This leads to choosing m to minimise the **Akaike** information criteria (AIC)

$$\mathrm{AIC}_m = 2\left(d_m - \widehat{\ell}_m\right),$$

where the first takes $d_m = \dim(\theta_m)$.

- Profiling over many nuisance parameters can lead to completely wrong inferences.
- Even when the number of nuisance parameters is $o(n)$ we may run into trouble: in general

$$\text{Bias } (\widehat{\psi}; \psi) = O(d^3/n),$$

so for the bias to tend to zero in large samples we require $d = o(n^{1/3})$ for consistency of $\widehat{\psi}$. Hence bias increases with $\dim(\lambda)$, at least in general.

Find the profile log likelihood for σ^2 when $(y_{j1}, y_{j2}) \overset{\text{iid}}{\sim} \mathcal{N}(\mu_j, \sigma^2)$, for $j = 1, \ldots, n$.

- The overall log likelihood is

$$\ell(\sigma^2, \mu_1, \ldots, \mu_n) \equiv -\tfrac{1}{2}\left[(2n)\log\sigma^2 + \frac{1}{\sigma^2}\sum_{j=1}^{n}\left\{(y_{j1}-\mu_j)^2 + (y_{j2}-\mu_j)^2\right\}\right],$$

  and differentiation with respect to μ_j gives that $\widehat{\mu}_j = (y_{j1} + y_{j2})/2$, so as

$$\{a - (a+b)/2\}^2 + \{b - (a+b)/2\}^2 = (a-b)^2/2,$$

  we obtain

$$\ell_{\mathrm{p}}(\sigma^2) = -n\log\sigma^2 - \frac{1}{4\sigma^2}\sum_{j=1}^{n}(y_{j1}-y_{j2})^2, \quad \sigma^2 > 0.$$

- This is maximised at $\widehat{\sigma}_p^2 = (4n)^{-1}\sum_{j=1}^{n}(y_{j1}-y_{j2})^2$, but as $Y_{j1} - Y_{j2} \overset{\text{iid}}{\sim} \mathcal{N}(0, 2\sigma^2)$, we see that $\sigma_p^2 \xrightarrow{P} \sigma^2/2$ as $n \to \infty$; this is a completely inconsistent estimator. Hence the profile log likelihood has its asymptotic maximum in completely the wrong place (there are $d = n + 1$ parameters of which n are nuisance parameters).

How can we rescue 'ordinary' likelihood inference when there are many nuisance parameters?

Approaches to dealing with high-dimensional λ include:

- basing inference on a **marginal likelihood** or a **conditional likelihood**,

$$f(y;\psi,\lambda) = f(w;\psi) \times f(y \mid w;\psi,\lambda) = f(y \mid w_\psi;\psi) \times f(w_\psi;\psi,\lambda),$$

  where w_ψ may not depend on ψ — OK for any configuration of λs, but may lose information on ψ;
- constructing a **partial likelihood** (like the above, but harder to build);
- **higher-order inference**, via, e.g., a **modified profile likelihood**, which can approximate both conditional and marginal likelihoods;
- using **orthogonal parameters**, i.e., mapping $\lambda \mapsto \zeta(\lambda,\psi)$ which is orthogonal to ψ;
- using a **composite likelihood** in which λ does not appear; or
- taking $\lambda \sim h(\cdot)$ and using the **integrated likelihood** $\int f(y;\psi,\lambda)h(\lambda)\,\mathrm{d}\lambda$ — depends on h, like Bayesian inference.

Next we sketch some of the approaches.

- Replace profile log likelihood $\ell_{\mathrm{p}}(\psi)$ by the **modified profile log likelihood**

$$\ell_{\mathrm{mp}}(\psi) = \ell_{\mathrm{p}}(\psi) + m(\psi),$$

  with $m(\psi)$ chosen to make ℓ_{p} closer to a marginal or conditional log likelihood.
- Taking

$$m(\psi) = -\tfrac{1}{2} \log \left| \jmath_{\lambda\lambda}(\psi, \widehat{\lambda}_\psi) \right| + \log \left| \frac{\partial \widehat{\lambda}}{\partial \widehat{\lambda}_\psi^{\mathrm{T}}} \right|$$

  does this in some generality.
- The first term of $m(\psi)$ can be obtained numerically if need be, but
- the second term, a Jacobian needed to make ℓ_{mp} invariant to interest-preserving reparametrisation, is hard to compute in general.

- If $\widehat{\psi}$ is asymptotically independent of $\widehat{\lambda}$, we can hope that the effect on $\widehat{\psi}$ of estimating λ will be limited. If so, we say that ψ and λ are **orthogonal**.
- To see the effect of this, we expand the equation defining $\widehat{\lambda}_\psi$ around $\widehat{\theta}$, giving

$$\begin{aligned} 0 &= \frac{\partial \ell(\widehat{\theta}_\psi)}{\partial \lambda} = \frac{\partial \ell(\widehat{\theta})}{\partial \lambda} + \frac{\partial^2 \ell(\widehat{\theta})}{\partial \lambda \partial \theta^{\mathrm{T}}}(\widehat{\theta}_\psi - \widehat{\theta}) + \cdots \\ &= \frac{\partial^2 \ell(\widehat{\theta})}{\partial \lambda \partial \lambda^{\mathrm{T}}}(\widehat{\lambda}_\psi - \widehat{\lambda}) + \frac{\partial^2 \ell(\widehat{\theta})}{\partial \lambda \partial \psi^{\mathrm{T}}}(\psi - \widehat{\psi}) + \cdots \\ &= \widehat{\jmath}_{\lambda\lambda}(\widehat{\lambda}_\psi - \widehat{\lambda}) + \widehat{\jmath}_{\lambda\psi}(\psi - \widehat{\psi}) + \cdots \end{aligned}$$

  which implies that

$$\widehat{\lambda}_\psi = \widehat{\lambda} + \widehat{\jmath}_{\lambda\lambda}^{-1}\widehat{\jmath}_{\lambda\psi}(\widehat{\psi} - \psi) + \cdots$$

- Hence if we can arrange the model so that $\widehat{\jmath}_{\lambda\psi} \approx 0$, then $\widehat{\lambda}_\psi$ will depend only weakly on ψ, and we can ignore the Jacobian term in the modified profile likelihood.
- This suggests mapping an original parametrisation (ψ, γ) to (ψ, λ), where $\lambda = \lambda(\psi, \gamma)$ is orthogonal to ψ.

- Writing $\gamma = \gamma(\psi, \lambda)$ gives
$$\ell(\psi, \lambda) = \ell^* \{\psi, \gamma(\psi, \lambda)\},$$
and differentiation with respect to ψ and λ leads to
$$\frac{\partial^2 \ell}{\partial \lambda \partial \psi} = \frac{\partial \gamma^{\mathrm{T}}}{\partial \lambda} \frac{\partial^2 \ell^*}{\partial \gamma \partial \psi} + \frac{\partial \gamma^{\mathrm{T}}}{\partial \lambda} \frac{\partial^2 \ell^*}{\partial \gamma \partial \gamma^{\mathrm{T}}} \frac{\partial \gamma}{\partial \psi} + \frac{\partial^2 \gamma^{\mathrm{T}}}{\partial \lambda \partial \psi} \frac{\partial \ell^*}{\partial \gamma}.$$
- For orthogonality this must have expectation zero, so
$$0 = \frac{\partial \gamma^{\mathrm{T}}}{\partial \lambda} i^*_{\gamma\psi} + \frac{\partial \gamma^{\mathrm{T}}}{\partial \lambda} i^*_{\gamma\gamma} \frac{\partial \gamma}{\partial \psi},$$
where $i^*_{\gamma\psi}$ and $i^*_{\gamma\gamma}$ are components of the expected information matrix in the non-orthogonal parametrization, so λ solves the system of q PDEs
$$\frac{\partial \gamma}{\partial \psi} = -i^{*-1}_{\gamma\gamma}(\psi, \gamma) i^*_{\gamma\psi}(\psi, \gamma).$$
- In fact an explicit expression for λ in terms of ψ and γ is not needed to compute ℓ_{mp} in the new parametrisation.

- A solution (possibly numerical) always exists when $\dim(\psi) = 1$, but need not exist when ψ is vector, because then we must simultaneously solve

$$\frac{\partial \gamma}{\partial \psi_1} = -i^{*-1}_{\gamma\gamma}(\psi, \gamma) i^{*}_{\gamma\psi_1}(\psi, \gamma), \quad \frac{\partial \gamma}{\partial \psi_2} = -i^{*-1}_{\gamma\gamma}(\psi, \gamma) i^{*}_{\gamma\psi_2}(\psi, \gamma),$$

  for all γ, ψ_1 and ψ_2, but the compatibility condition

$$\frac{\partial^2 \gamma}{\partial \psi_1 \partial \psi_2} = \frac{\partial^2 \gamma}{\partial \psi_2 \partial \psi_1}$$

  may fail.

- Used when full likelihood can't be computed but densities for distinct subsets of the observations, $y_{\mathcal{S}_1}, \ldots, y_{\mathcal{S}_C}$, are available, can use a **composite (log) likelihood**

$$\ell_{\mathrm{C}}(\theta) = \sum_{c=1}^{C} \log f(y_{\mathcal{S}_c}; \theta).$$

- The choice of subsets $\mathcal{S}_1, \ldots, \mathcal{S}_C$ determines what parameters can be estimated.
- Special cases:
  - **independence likelihood** takes $\mathcal{S}_j = \{y_j\}$ and treats (possibly dependent) y_j as independent;
  - **pairwise likelihood** uses subsets of distinct pairs $\{y_j, y_{j'}\}$.
- May be useful with spatial data, and then contributions from distant pairs may be downweighted or dropped entirely.
- $\ell_{\mathrm{C}}(\theta)$ satisfies the first Bartlett identity, so can give consistent estimators $\widehat{\theta}$, but requires a sandwich variance matrix (or some other approach) to estimate $\mathrm{var}(\widehat{\theta})$.
- Model comparisons use the **composite likelihood information criterion**

$$\mathrm{CLIC} = 2\left[b(\widehat{\theta}) - \ell_{\mathrm{C}}(\widehat{\theta})\right].$$

- Empirical likelihood allows nonparametric estimation of constrained distributions.
- If $y_1, \ldots, y_n \overset{\text{iid}}{\sim} G$, then the $\widehat{G}$ that maximises the 'nonparametric likelihood'

$$L(G) = \prod_{j=1}^{n} G(\mathrm{d}y_j) = \prod_{j=1}^{n} p_j, \quad \text{subject to} \quad p_j \geq 0, \quad \sum_{j=1}^{n} p_j \leq 1,$$

  sets $G(\mathrm{d}y_j) = \widehat{p}_j \equiv n^{-1}$: $\widehat{G}$ is the empirical distribution function of $y_1, \ldots, y_n$.
- Adding the constraint $\mathrm{E}\{c(Y;\theta)\} = 0$ leads to maximising

$$\sum_{j=1}^{n} \log p_j \quad \text{subject to} \quad p_j \geq 0, \quad \sum_{j=1}^{n} p_j \leq 1, \quad \sum_{j=1}^{n} p_j c(y_j;\theta) = 0.$$

  Solving this with Lagrange multipliers shows that we must find $a = a_\theta$ to solve

$$\sum_{j=1}^{n} \frac{c(y_j;\theta)}{n\{1 + ac(y_j;\theta)\}} = 0$$

  giving $\widehat{a} = \widehat{a}_\theta$, $\widehat{p}_j(\theta) = n^{-1}/\{1 + \widehat{a}c(y_j;\theta)\}$ and empirical likelihood ratio statistic $w(\theta) = 2\sum \log\{1 + \widehat{a}c(y_j;\theta)\}$
- The usual χ^2 result applies to $w(\theta)$.

Comments

- Other likelihoods and/or likelihood-like functions are widely used, especially
  - **partial likelihood**, used to eliminate nuisance functions for inference (survival data),
  - **quasi-likelihood**, used to model over-dispersion in exponential family models,
  - **pseudo-likelihood**, treats data as Gaussian even when they are not (econometrics), and
- Strengths of likelihood approach:
  - heuristic as plausibility of a model as explanation of data;
  - we 'just' have to write down the density of the observed data;
  - invariance to data and parameter transformations;
  - general (and 'optimal') approximate theory for inference in regular models;
  - close links to Bayesian inference.
- Weaknesses of likelihood approach:
  - requires 'parametric' model for data;
  - can fail in high-dimensional settings;
  - not all models are regular.