3 Likelihood Theory

slide 92

3.1 Basic Results

slide 93

Motivation

- ☐ Likelihood
  - 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!
- ☐ Plan below:
  - give (fairly) general setup;
  - prove main results for scalar parameter;
  - discussion of inference;
  - vector parameter, nuisance parameters, ...

Basic setup

- ☐ 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,

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

  where the inequality holds because $-\log x$ is convex and is strict unless $f \equiv g$ (Jensen).
- ☐ 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 $n \to \infty$.

Regular models

- Notation: $\nabla h(\theta) = \partial h(\theta)/\partial\theta$ and $\nabla^2 h(\theta) = \nabla\nabla^{\mathrm{T}} h(\theta) = \partial^2 h(\theta)/\partial\theta\partial\theta^{\mathrm{T}}$.
- The asymptotic properties of the MLE rely on **regularity conditions**:

(C1) θ_g is unique and interior to $\Theta \subset \mathbb{R}^d$ for some finite d, and Θ is compact;

(C2) the densities f_θ defined by any two different values of $\theta \in \Theta$ are distinct;

(C3) there is a neighbourhood $\mathcal{N}$ of θ_g within which the first three derivatives of the log likelihood with respect to θ exist almost surely, and for $r, s, t = 1, \ldots, d$ satisfy $|\partial^3 \log f(Y;\theta)/\partial\theta_r\partial\theta_s\partial\theta_t| < m(Y)$ with $\mathrm{E}_g\{m(Y)\} < \infty$; and

(C4) within $\mathcal{N}$, the $d \times d$ matrices

$$\imath_1(\theta) = \mathrm{E}_g\left\{-\nabla^2 \log f(Y;\theta)\right\}, \quad \hbar_1(\theta) = \mathrm{E}_g\left\{\nabla \log f(Y;\theta)\nabla^{\mathrm{T}} \log f(Y;\theta)\right\},$$

are finite and positive definite. When $g = f_{\theta_g}$ we shall see that $\hbar_1(\theta_g) = \imath_1(\theta_g)$.

Regularity conditions

- (C1) ensures that $\widehat{\theta}$ can be 'on all sides' of θ_g in the limit — if it fails, then any limiting distribution cannot be normal;
- (C2) is essential for consistency, otherwise $\widehat{\theta}$ might not converge — it often fails in mixture models, for which care is needed;
- (C3) is needed to bound terms of a Taylor series — can be replaced by other conditions, see van der Vaart (1998, *Asymptotic Statistics*, Chapter 5); and
- (C4) ensures that the asymptotic variance of $\widehat{\theta}$ is positive definite.

Consistency of the MLE

Lemma 49 *If $Y_1, \ldots, Y_n \sim g$ and $n \to \infty$, then under (C1) and (C2) a sequence of maximum likelihood estimators $\widehat{\theta}$ exists such that $\widehat{\theta} \xrightarrow{P} \theta_g$.*

This result:

- does not require f_θ to be smooth, so it is quite general;
- guarantees that a consistent sequence exists, but not that we can find it;
- but if the log likelihood is concave (as in exponential families, for example), then there is (at most) one maximum for any n, and if it exists this must converge to θ_g;
- can be generalized to vector θ, but the argument is more delicate.

Note to Lemma 49

- ☐ We prove this for θ scalar.
- ☐ As the θs correspond to different densities, precisely one θ_g minimises $\mathrm{KL}(g, f_\theta)$.
- ☐ Take any $\varepsilon > 0$ and let $\theta_+, \theta_- = \theta_g \pm \varepsilon$, write $D_n(\theta) = \overline{\ell}(\theta_g) - \overline{\ell}(\theta)$, so $D_n(\theta_g) = 0$, and note that as $n \to \infty$,

  $$D_n(\theta_+) \xrightarrow{P} \mathrm{KL}(g, f_{\theta_+}) - \mathrm{KL}(g, f_{\theta_g}) = a_+ > 0, \quad D_n(\theta_-) \xrightarrow{P} \mathrm{KL}(g, f_{\theta_-}) - \mathrm{KL}(g, f_{\theta_g}) = a_- > 0.$$

- ☐ If A_n and B_n denote the events $D_n(\theta_+) > 0$ and $D_n(\theta_-) > 0$, Boole's inequality gives

  $$\mathrm{P}(A_n \cap B_n) = 1 - \mathrm{P}(A_n^c \cup B_n^c) \geq 1 - \mathrm{P}(A_n^c) - \mathrm{P}(B_n^c).$$

  Now

  $$\mathrm{P}(A_n^c) = \mathrm{P}\{D_n(\theta_+) \leq 0\} = \mathrm{P}\{a_+ - D_n(\theta_+) \geq a_+\} \leq \mathrm{P}\{|D_n(\theta_+) - a_+| \geq a_+\} \to 0, \quad n \to \infty,$$

  and likewise $\mathrm{P}(B_n^c) \to 0$. Hence $\mathrm{P}(A_n \cap B_n) \to 1$.
- ☐ Hence there is a local minimum of $D_n(\theta)$, or equivalently a local maximum of $\overline{\ell}(\theta)$, inside the interval $(\theta_g - \epsilon, \theta_g + \epsilon)$ with probability one as $n \to \infty$, and as this is true for arbitrary ε, the corresponding sequence of maximisers $\widehat{\theta}$ satisfies $\mathrm{P}(|\widehat{\theta} - \theta_g| > \varepsilon) \to 0$ and therefore is consistent.

Asymptotic normality of the MLE

Theorem 50 *If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g$, then under (C1)–(C4) the consistent sequence of maximum likelihood estimators $\widehat{\theta}$ satisfies*

$$n^{1/2}(\widehat{\theta} - \theta_g) \xrightarrow{D} \mathcal{N}_d\{0, \imath_1^{-1}(\theta_g)\hbar_1(\theta_g)\imath_1^{-1}(\theta_g)\}, \quad n \to \infty,$$

where for a single observation Y we define

$$\imath_1(\theta) = \mathrm{E}_g\left\{-\nabla^2 \log f(Y;\theta)\right\}, \quad \hbar_1(\theta) = \mathrm{E}_g\left\{\nabla \log f(Y;\theta) \nabla^{\mathrm{T}} \log f(Y;\theta)\right\}.$$

- ☐ This implies that for large n we can use the approximation

  $$\widehat{\theta} \,\dot{\sim}\, \mathcal{N}_d\{\theta_g, \imath^{-1}(\theta_g)\hbar(\theta_g)\imath^{-1}(\theta_g)\},$$

  where $\imath(\theta) = n\imath_1(\theta)$ and $\hbar(\theta) = n\hbar_1(\theta)$ correspond to a random sample of size n.
- ☐ This provides tests and confidence intervals based on the approximate pivots

  $$v_{rr}^{-1/2}(\widehat{\theta}_r - \theta_{g,r}) \,\dot{\sim}\, \mathcal{N}(0,1), \quad r = 1, \ldots, d,$$

  where v_{rr} are the diagonal elements of an estimate of $\imath^{-1}(\theta_g)\hbar(\theta_g)\imath^{-1}(\theta_g)$.
- ☐ When $g = f_{\theta_g}$, $\imath_1(\theta_g) = \hbar_1(\theta_g)$ and the variance (matrix) becomes $\imath(\theta_g)^{-1}$.

Note to Theorem 50: A (fairly) simple argument

☐ Write

$$0 = \nabla\overline{\ell}(\widehat{\theta}) = \nabla\overline{\ell}(\theta_g) + \int_0^1 \nabla^2\overline{\ell}\{\theta_g + t(\widehat{\theta} - \theta_g)\}(\widehat{\theta} - \theta_g)\,\mathrm{d}t,$$

and note that $U_n = n^{1/2}\nabla\overline{\ell}(\theta_g) \xrightarrow{D} U \sim \mathcal{N}_d\{0, \hbar_1(\theta_g)\}$, so writing $Z_n = n^{1/2}(\widehat{\theta} - \theta_g)$ we have

$$\imath_1(\theta_g)^{-1}U_n = \imath_1(\theta_g)^{-1}\left\{-\int_0^1 \nabla^2\overline{\ell}(\theta_g + tn^{-1/2}Z_n)\,\mathrm{d}t\right\}Z_n = \imath_1(\theta_g)^{-1}J_n^*Z_n,$$

say, and as $n \to \infty$, $J_n^* \doteq -\int_0^1 \nabla^2\overline{\ell}(\theta_g)\,\mathrm{d}t \xrightarrow{P} \imath_1(\theta_g)$ and thus $\imath_1(\theta_g)^{-1}J_n^* \xrightarrow{P} I_d$. Hence

$$\imath_1(\theta_g)^{-1}J_n^*Z_n = \imath_1(\theta_g)^{-1}U_n \xrightarrow{D} \imath_1(\theta_g)^{-1}U \sim \mathcal{N}_d\{0, \imath_1(\theta_g)^{-1}\hbar_1(\theta_g)\imath_1(\theta_g)^{-1}\}.$$

☐ For a more careful treatment of the integral, we need a *uniform law of large numbers (ULLN)*, which requires that $J_n(\theta) = -\nabla^2\overline{\ell}(\theta)$ is measurable and continuous in θ within a compact subset $\mathcal{N}'$ of $\mathcal{N}$, for almost all y, and that there exists a function $d(Y)$ whose expectation is finite and for which $\|J_n(\theta)\| < d(Y)$ for all $\theta \in \mathcal{N}'$, where $\|\cdot\|$ is a matrix norm. Then $\mathrm{E}\{J_n(\theta)\} = \imath_1(\theta)$ is continuous in θ and

$$\sup_{\theta\in\mathcal{N}} \|J_n(\theta) - \imath_1(\theta)\| \xrightarrow{P} 0, \quad n \to \infty.$$

☐ Let $\delta > 0$ be small enough that $B_\delta = \{\theta : |\theta - \theta_g| \le \delta\} \subset \mathcal{N}'$ and let $A_n = \{|n^{-1/2}Z_n| \le \delta\}$ and $C_n = \|J_n^* - \imath_1(\theta_g)\|$. Then for $\varepsilon > 0$ we have

$$\mathrm{P}(C_n > \varepsilon) = \mathrm{P}(\{C_n > \varepsilon\} \cap A_n) + \mathrm{P}(\{C_n > \varepsilon\} \cap A_n^c) \le \mathrm{P}(\{C_n > \varepsilon\} \cap A_n) + \mathrm{P}(A_n^c),$$

where the last term tends to zero because $n^{-1/2}Z_n = \widehat{\theta} - \theta_g \xrightarrow{P} 0$. Now if A_n holds, then $\theta_g + tn^{-1/2}Z_n \in B_\delta$ when $0 \le t \le 1$, so

$$\begin{aligned}
C_n &= \left\|\int_0^1 \left\{J_n(\theta_g + tn^{-1/2}Z_n) - \imath_1(\theta_g + tn^{-1/2}Z_n) + \imath_1(\theta_g + tn^{-1/2}Z_n) - \imath_1(\theta_g)\right\}\,\mathrm{d}t\right\| \\
&\le \int_0^1 \left\|J_n(\theta_g + tn^{-1/2}Z_n) - \imath_1(\theta_g + tn^{-1/2}Z_n)\right\|\,\mathrm{d}t + \int_0^1 \left\|\imath_1(\theta_g + tn^{-1/2}Z_n) - \imath_1(\theta_g)\right\|\,\mathrm{d}t \\
&\le \sup_{\theta\in B_\delta} \|J_n(\theta) - \imath_1(\theta)\| + \sup_{\theta\in B_\delta} \|\imath_1(\theta) - \imath_1(\theta_g)\| \\
&= D_n + E_n,
\end{aligned}$$

say. If $C_n > \varepsilon$ then at least one of D_n and E_n must exceed $\varepsilon/2$, so

$$\begin{aligned}
\mathrm{P}(\{C_n > \varepsilon\} \cap A_n) &\le \mathrm{P}\left(\{\{D_n > \varepsilon/2\} \cup \{E_n \ge \varepsilon/2\}\} \cap A_n\right) \\
&\le \mathrm{P}(\{D_n > \varepsilon/2\} \cap A_n) + \mathrm{P}(\{E_n \ge \varepsilon/2\} \cap A_n) \\
&\le \mathrm{P}(D_n > \varepsilon/2) + \mathrm{P}(E_n > \varepsilon/2).
\end{aligned}$$

Now $D_n \xrightarrow{P} 0$ using the ULLN, and the continuity of $\imath_1(\theta)$ at θ_g implies that E_n can be made smaller than $\varepsilon/2$ by a suitable choice of $\delta > 0$, in which case

$$\begin{aligned}
\mathrm{P}(C_n > \varepsilon) &\le \mathrm{P}(\{C_n > \varepsilon\} \cap A_n) + \mathrm{P}(A_n^c) \\
&\le \mathrm{P}\left(D_n > \varepsilon/2\right) + \mathrm{P}\left(E_n > \varepsilon/2\right) + \mathrm{P}(A_n^c) \\
&\to 0, \quad n \to \infty,
\end{aligned}$$

which implies that $J_n^* \xrightarrow{P} \imath_1(\theta_g)$ and therefore that $\imath_1(\theta_g)^{-1}J_n^* \xrightarrow{P} I_d$, as required.

Note to Theorem 50: Another approach

☐ We first note that under the given conditions, θ_g gives a stationary point of $\mathrm{KL}(g, f_\theta)$, and therefore

$$0 = \nabla \mathrm{KL}(g, f_\theta)|_{\theta=\theta_g} = -\nabla \int \log f(y;\theta) g(y)\,\mathrm{d}y \Big|_{\theta=\theta_g} = -\int \nabla \log f(y;\theta)\Big|_{\theta=\theta_g} g(y)\,\mathrm{d}y,$$

so $\mathrm{E}_g\{\nabla \log f(Y;\theta)\} = 0$.

☐ As $\widehat{\theta}$ gives a local maximum of the differentiable function $\overline{\ell}(\theta) = n^{-1}\sum_{j=1}^n \log f(Y_j;\theta)$,

$$0 = \nabla \overline{\ell}(\widehat{\theta}) = n^{-1}\sum_{j=1}^n \nabla \log f(Y_j;\widehat{\theta}),$$

and (supposing now that θ is scalar, to simplify the expressions), Taylor series expansion gives

$$0 = \nabla \overline{\ell}(\theta_g) + (\widehat{\theta}-\theta_g)\nabla^2\overline{\ell}(\theta_g) + \tfrac{1}{2}(\widehat{\theta}-\theta_g)^2\nabla^3\overline{\ell}(\theta^*),$$

where θ^* lies between θ_g and $\widehat{\theta}$ (so $\theta^* \xrightarrow{P} \theta_g$). Hence

$$n^{1/2}(\widehat{\theta}-\theta_g) = \frac{n^{1/2}\nabla\overline{\ell}(\theta_g)}{-\nabla^2\overline{\ell}(\theta_g) - R_n/2}, \quad R_n = (\widehat{\theta}-\theta_g)\nabla^3\overline{\ell}(\theta^*). \tag{3}$$

☐ Now

$$n^{1/2}\nabla\overline{\ell}(\theta_g) = n^{-1/2}\sum_{j=1}^n \nabla \log f(Y_j;\theta_g)$$

has mean (vector) zero and variance (matrix)

$$\mathrm{var}\left\{n^{-1/2}\sum_{j=1}^n \nabla \log f(Y_j;\theta_g)\right\} = n^{-1}\sum_{j=1}^n \mathrm{E}_g\{\nabla \log f(Y_j;\theta_g)\nabla^{\mathrm{T}} \log f(Y_j;\theta_g)\} = \hbar_1(\theta_g).$$

so the numerator of (3) converges in distribution to $\mathcal{N}\{0, \hbar_1(\theta_g)\}$, using the CLT.

☐ Moreover the weak law of large numbers gives

$$-\nabla^2\overline{\ell}(\theta_g) = -\frac{1}{n}\sum_{j=1}^n \nabla^2 \log f(Y_j;\theta_g) \xrightarrow{P} \imath_1(\theta_g).$$

☐ Lemma 51 shows that $R_n \xrightarrow{P} 0$, so the denominator of (3) tends in probability to $\imath_1(\theta_g)$.

☐ Putting the pieces together, we find that

$$n^{1/2}(\widehat{\theta}-\theta_g) \xrightarrow{D} \mathcal{N}_d\{0, \imath_1(\theta_g)^{-1}\hbar_1(\theta_g)\imath_1(\theta_g)^{-1}\}, \quad n\to\infty,$$

where the variance formula is also valid when $\imath_1$ and $\hbar_1$ are $d\times d$ matrices.

☐ The information quantities based on a random sample of size n are $\imath(\theta_g) = n\imath_1(\theta_g)$ and $\hbar(\theta_g) = n\hbar_1(\theta_g)$, giving

$$\widehat{\theta} \dot{\sim} \mathcal{N}_d(\theta_g, \imath(\theta_g)^{-1}\hbar(\theta_g)\imath(\theta_g)^{-1}\},$$

in which the variance is of the usual order $1/n$.

Note: A useful lemma

Lemma 51 *Under the conditions of Theorem 50,* $R_n = (\widehat{\theta} - \theta_g)\nabla^3\overline{\ell}(\theta^*) \xrightarrow{P} 0$ *as* $n \to \infty$.

- For $\varepsilon > 0$, $B_n = \{|R_n| > \varepsilon\}$, $A_n = \{|\widehat{\theta} - \theta_g| > \delta\}$ and $\delta > 0$ small enough that $\mathcal{N}$ contains a ball of radius δ around θ_g, we have

$$\mathrm{P}(|R_n| > \varepsilon) = \mathrm{P}(B_n \cap A_n) + \mathrm{P}(B_n \cap A_n^c) \leq \mathrm{P}(A_n) + \mathrm{P}(B_n \cap A_n^c),$$

  where the first term tends to zero because the sequence $\widehat{\theta}$ is consistent.
- If $|\widehat{\theta} - \theta_g| < \delta$, then (C3) implies that

$$|R_n| \leq \delta n^{-1} \sum_{j=1}^{n} |\partial^3 \log f(Y_j; \theta^*)/\partial\theta^3| \leq \delta n^{-1} \sum_{j=1}^{n} m(Y_j) = \delta\overline{M}_n,$$

  say, and clearly $\overline{M}_n \xrightarrow{P} M$, say. Therefore

$$\mathrm{P}(B_n \cap A_n^c) = \mathrm{P}(B_n \cap |\widehat{\theta} - \theta_g| > \delta) \leq \mathrm{P}(B_n \cap |R_n| \leq \delta\overline{M}_n)$$

  and for $\eta > 0$ this equals

$$\mathrm{P}(B_n \cap |R_n| \leq \delta\overline{M}_n \cap \overline{M}_n \leq M + \eta) + \mathrm{P}(B_n \cap |R_n| \leq \delta\overline{M}_n \cap \overline{M}_n > M + \eta),$$

  which is bounded by

$$\mathrm{P}\{|R_n| > \varepsilon \cap |R_n| \leq \delta(M + \eta)\} + \mathrm{P}(|\overline{M}_n - M| > \eta).$$

  The last term here tends to zero, because $\overline{M}_n \xrightarrow{P} M$, and the first can be made equal to zero by choosing δ such that $\delta(M + \eta) < \varepsilon$. This proves the lemma.

Classical asymptotics

- The true model is supposed to lie in the candidate family, i.e., $g \in \mathcal{F}$, so $\theta_g \in \Theta$.
- We saw on slide 38 that the moments of the $d \times 1$ **score vector** $U(\theta) = \nabla\ell(\theta)$ are given under mild conditions by the Bartlett identities, i.e.,

$$\mathrm{E}\{U(\theta)\} = 0, \quad \mathrm{var}\{U(\theta)\} = \mathrm{E}\left\{\nabla\ell(\theta)\nabla^{\mathrm{T}}\ell(\theta)\right\} = \mathrm{E}\left\{-\nabla^2\ell(\theta)\right\}, \quad \ldots$$

- Hence $\imath(\theta) = \hbar(\theta)$, and $\imath(\theta) = n\imath_1(\theta) = n\hbar_1(\theta)$ when $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f_{\theta_g}$.
- 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 ψ 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.

Note: Stable interpretation of a parameter

☐ To put some mathematical flesh on the discussion, suppose that $g(y) = f(y; \theta, \gamma)$ and the assumed model is $f(y; \theta, 0)$. Then for small γ, $\theta_g \equiv \theta_\gamma$ satisfies

$$\begin{aligned} 0 &= \int \nabla_\theta \log f(y; \theta_\gamma, 0) f(y; \theta, \gamma) \, \mathrm{d}y \\ &= \int \left\{ \nabla_\theta \log f(y; \theta, \gamma) + \nabla_\theta^2 \log f(y; \theta, \gamma)(\theta_\gamma - \theta) + \nabla_\gamma^{\mathrm{T}} \nabla_\theta \log f(y; \theta, \gamma)(0 - \gamma) + \cdots \right\} f(y; \theta, \gamma) \, \mathrm{d}y \\ &= 0 - \imath_{\theta\theta}(\theta, \gamma)(\theta_\gamma - \theta) + \imath_{\theta\gamma}(\theta, \gamma)\gamma + o(\gamma), \end{aligned}$$

which implies that the effect of incorrectly assuming that $\gamma = 0$ is that $\widehat{\theta}$ converges to

$$\theta_\gamma = \theta + \imath_{\theta\theta}^{-1}(\theta, \gamma)\imath_{\theta\gamma}(\theta, \gamma)\gamma + o(\gamma).$$

☐ It is also easy to check that $\hbar_{\theta\theta}(\theta, 0) = \imath_{\theta\theta}(\theta, 0) + O(\gamma)$, so the two matrices become equal if $\gamma \to 0$, in which case $\imath_1(\theta, \gamma)^{-1} \hbar_1(\theta, \gamma) \imath_1(\theta, \gamma)^{-1} \to \imath_1(\theta, \gamma)^{-1}$, which implies that for small γ we have

$$n^{1/2}(\widehat{\theta} - \theta) = n^{1/2}(\widehat{\theta} - \theta_\gamma) + n^{1/2}(\theta_\gamma - \theta) \dot{\sim} \mathcal{N}_d\{0, \imath_1(\theta, \gamma)^{-1}\} + n^{1/2}(\theta_\gamma - \theta).$$

☐ Now if $\gamma = n^{-a}\delta$ for some $a > 0$, then

$$n^{1/2}(\theta_\gamma - \theta) = n^{1/2-a} \imath_{\theta\theta}^{-1}(\theta, \gamma)\imath_{\theta\gamma}(\theta, \gamma)\delta,$$

which will tend to infinity if $a < 1/2$ (should be obvious asymptotically), to zero if $a > 1/2$ (can be ignored asymptotically) and to a constant if $a = 1/2$. Hence there is an asymptotic bias for $\widehat{\theta}$ if there is misspecification, $\delta \neq 0$, unless $\imath_{\theta\gamma}(\theta, \gamma) = 0$, i.e., the information matrix covariance for the scores for θ and γ is zero. This is known as orthogonality of θ and γ; see later.

In practice ...

- We usually assume classical asymptotics and replace the sandwich matrix $\imath(\theta_g)^{-1}\hbar(\theta_g)\imath(\theta_g)^{-1}$ 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 $\widehat{\jmath}^{-1}$.
- If we must make the sandwich we can replace $\imath(\theta_g)$ by $\widehat{\jmath}$ and $\hbar(\theta_g)$ by (e.g.)

$$\widehat{\hbar} = \sum_{j=1}^{n} \nabla \log f(Y_j; \widehat{\theta}) \nabla^{\mathrm{T}} \log f(Y_j; \widehat{\theta}),$$

  though $\widehat{\jmath}^{-1}\widehat{\hbar}\widehat{\jmath}^{-1}$ can be unstable because $\widehat{\hbar}$ misbehaves.

Related statistics

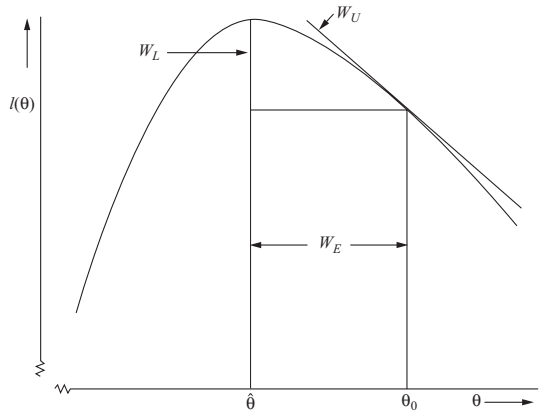


Figure 6.2. Three asymptotically equivalent ways, all based on the log likelihood function of testing null hypothesis $\theta = \theta_0$: W_E, horizontal distance; W_L vertical distance; W_U slope at null point.

From Cox (2006, *Principles of Statistical Inference*)

Related statistics

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

$$\begin{aligned} T = t(\theta_g) &= \widehat{\jmath}^{1/2}(\widehat{\theta} - \theta_g) \dot{\sim} \mathcal{N}(0,1), && \text{Wald statistic}, \\ S = s(\theta_g) &= \widehat{\jmath}^{-1/2} U(\theta_g) \dot{\sim} \mathcal{N}(0,1), && \text{score statistic}, \\ W = w(\theta_g) &= 2\{\ell(\widehat{\theta}) - \ell(\theta_g)\} \dot{\sim} \chi^2_1, && \text{likelihood ratio statistic}, \\ R = r(\theta_g) &= \mathrm{sign}(\widehat{\theta} - \theta_g) w(\theta_g)^{1/2} \dot{\sim} \mathcal{N}(0,1), && \text{likelihood root}. \end{aligned}$$

The likelihood root has other names (e.g., directed likelihood ratio statistic).

☐ The distribution of W follows from the expansion on the previous slide.

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

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

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 or R take the shape of ℓ into account and are parametrisation-invariant;

☐ in small samples the distributional approximations for W and R 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 or R are unreliable);

☐ the main use of S is for testing in situations where maximisation of ℓ is awkward, and then $\widehat{\jmath}$ is often replaced by $\imath(\theta_g)$;

☐ a variant of R, the **modified likelihood root**

$$R^* = r^*(\theta_g) = r(\theta_g) + \frac{1}{r(\theta_g)} \log \frac{q(\theta_g)}{r(\theta_g)},$$

often gives almost perfect inferences even in small samples (more later ...).

Example 52 *Compute the above statistics when* $y_1, \ldots, y_n \overset{\text{iid}}{\sim} \exp(\theta)$ *and compare the resulting inferences with those from an exact pivot.*

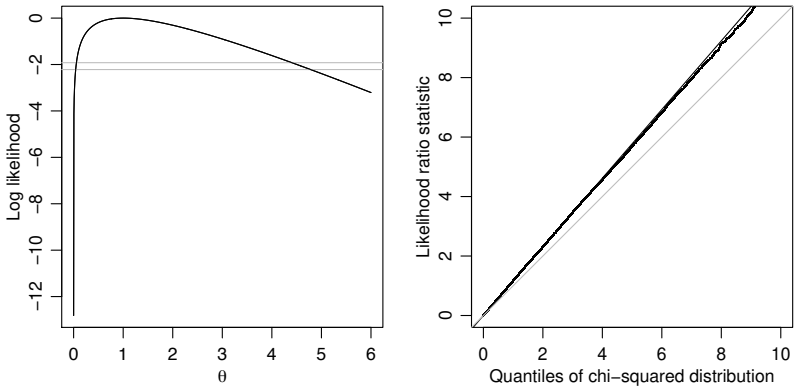
Note to Example 52

- ☐ The log likelihood is $\ell(\theta) = n(\log\theta - \theta\overline{y})$, for $\theta > 0$, which is clearly unimodal with $\widehat{\theta} = 1/\overline{y}$ and $\jmath(\theta) = n/\theta^2$.
- ☐ Hence

$$\begin{aligned} t(\theta) &= n^{1/2}(1-\theta\,\overline{y}), \\ s(\theta) &= n^{1/2}\{1/(\theta\,\overline{y}) - 1\}, \\ w(\theta) &= 2n\left\{\theta\,\overline{y} - \log(\theta\,\overline{y}) - 1\right\}, \\ r(\theta) &= \mathrm{sign}(1-\theta\,\overline{y})\left[2n\left\{\theta\,\overline{y} - \log(\theta\,\overline{y}) - 1\right\}\right]^{1/2}. \end{aligned}$$

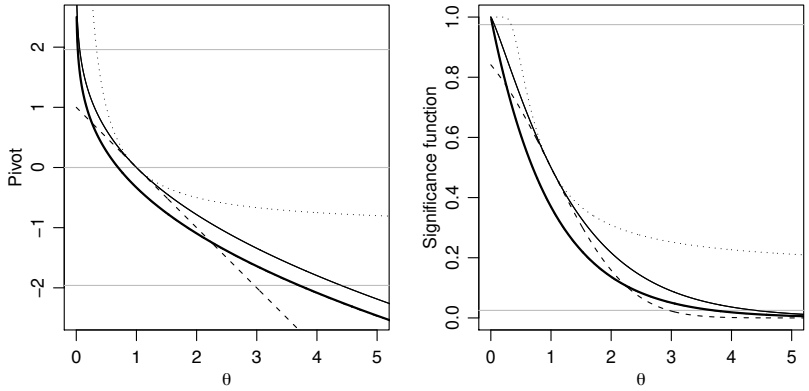
- ☐ The exact pivot is $\theta\sum Y_j$ whose distribution is gamma with unit scale and shape parameter n.
- ☐ Consider an exponential sample with $n = 1$ and $\overline{y} = 1$; then $\widehat{\jmath} = 1$. The log likelihood $\ell(\theta)$, shown in the left-hand panel of the figure, is unimodal but strikingly asymmetric, suggesting that confidence intervals based on an approximating normal distribution for $\widehat{\theta}$ will be poor. The right-hand panel is a chi-squared probability plot in which the ordered values of simulated $w(\theta)$ are graphed against quantiles of the χ^2_1 distribution—if the simulations lay along the diagonal line $x = y$, then this distribution would be a perfect fit. The simulations do follow a straight line rather closely, but with slope $(1 + b/n)\chi^2_1$, where $b = 0.1544$. This indicates that the distribution of the Bartlett-adjusted likelihood ratio statistic $w(\theta)/(1 + b/n)$ would be essentially χ^2_1. The 95% confidence intervals for θ based on the unadjusted and adjusted likelihood ratio statistics are $(0.058, 4.403)$ and $(0.042, 4.782)$ respectively.

Exponential example



Likelihood inference for exponential sample of size $n = 1$. Left: log likelihood $\ell(\theta)$. Intersection of the function with the two horizontal lines gives two 95% confidence intervals for θ: the upper line is based on the χ^2_1 approximation to the distribution of $w(\theta)$, and the lower line is based on the Bartlett-corrected statistic. Right: comparison of simulated values of likelihood ratio statistic $w(\theta)$ with χ^2_1 quantiles. The χ^2_1 approximation is shown by the line of unit slope, while the $(1 + b/n)\chi^2_1$ approximation is shown by the upper straight line.

Exponential example



Approximate pivots and P-values based on an exponential sample of size $n = 1$. Left: likelihood root $r(\theta)$ (solid), score pivot $s(\theta)$ (dots), Wald pivot $t(\theta)$ (dashes), modified likelihood root $r^*(\theta)$ (heavy), and exact pivot $\theta \sum y_j$ (dot-dash). The modified likelihood root is indistinguishable from the exact pivot. The horizontal lines are at $0, \pm 1.96$. Right: corresponding confidence functions, with horizontal lines at 0.025 and 0.975.

Non-regular models

- ☐ The regularity conditions (C1)–(C4) 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 53 *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* $\exp(1)$*. Discuss.*

Note to Example 53

☐ 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 $f(\theta;\theta) \neq 0$.

☐ Owing to the independence,

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

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

$$\mathrm{P}(M \leq x) = (x/\theta)^n, \quad 0 < x < \theta.$$

Now

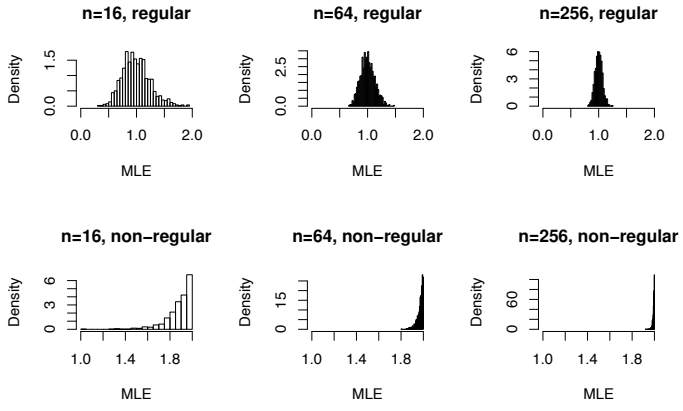
$$\mathrm{P}\left\{n(\theta - \widehat{\theta})/\theta \leq x\right\} = \mathrm{P}(\widehat{\theta} \geq \theta - x\theta/n) = 1 - \{(\theta - x\theta/n)/\theta\}^n \to 1 - \exp(-x),$$

as required. Note that:

- 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 limit 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 (Example 53, 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.



Vector case

- When θ is a vector and under classical asymptotics we base inference on the distributional approximations

$$\widehat{\theta} \dot{\sim} \mathcal{N}_d(\theta_g, \widehat{\jmath}^{-1}), \quad w(\theta_g) = 2\left\{\ell(\widehat{\theta}) - \ell(\theta_g)\right\} \dot{\sim} \chi^2_d, \quad s(\theta_g) = \widehat{\jmath}^{-1/2} U(\theta_g) \dot{\sim} \mathcal{N}_d(0, I_d),$$

  with
  - the first very commonly used for inferences on parameters;
  - the second used to test whether $\theta = \theta_g$;
  - the third much less used than the others, generally in the form $s(\theta_g)^{\mathrm{T}} s(\theta_g) \dot{\sim} \chi^2_d$.
- If θ divides into a $p \times 1$ **interest parameter** ψ and a $q \times 1$ **nuisance parameter** λ, then

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

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

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

Inference on ψ

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

$$\widehat{\psi} \dot{\sim} \mathcal{N}_p\left(\psi_g, \widehat{\jmath}^{\psi\psi}\right) \qquad \text{maximum likelihood estimator},$$
$$s(\psi_g) = \tilde{\ell}_\psi^{\mathrm{T}} \tilde{\jmath}^{\psi\psi} \tilde{\ell}_\psi \dot{\sim} \chi^2_p \qquad \text{score statistic},$$
$$w_{\mathrm{p}}(\psi_g) = 2\left\{\ell_{\mathrm{p}}(\widehat{\psi}) - \ell_{\mathrm{p}}(\psi_g)\right\} \dot{\sim} \chi^2_p \qquad \text{(generalized) likelihood ratio statistic},$$

  where we defined w_{p} using the **profile log likelihood** $\ell_{\mathrm{p}}(\psi) = \ell(\psi, \widehat{\lambda}_\psi) = \max_\lambda \ell(\psi, \lambda)$.
- If ψ is scalar ($p = 1$, the usual situation), the **likelihood root** is defined as

$$r(\psi_g) = \operatorname{sign}\left(\widehat{\psi} - \psi_g\right)\sqrt{w(\psi_g)} \dot{\sim} \mathcal{N}(0, 1).$$

- Properties:
  - inferences using $w(\psi_g)$ and $r(\psi_g)$ 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

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

Note: Large-sample distribution of the likelihood ratio statistic $w_{\mathrm{p}}(\psi_g)$

☐ We write

$$w_{\mathrm{p}}(\psi_g) = 2\{\ell(\widehat{\theta}) - \ell(\widehat{\theta}_\psi)\} = 2\{\ell(\widehat{\theta}) - \ell(\theta_g)\} - 2\{\ell(\widehat{\theta}_\psi) - \ell(\theta_g)\}$$

and use Taylor series to approximate both terms by quadratic forms in $\widehat{\theta} - \theta_g$ and $\widehat{\lambda}_\psi - \lambda_g$.

☐ To lighten the notation we let ℓ, $\tilde{\ell}$ and $\widehat{\ell}$ denote $\ell(\psi_g, \lambda_g)$, $\ell(\psi_g, \widehat{\lambda}_{\psi_g})$ and $\ell(\widehat{\psi}, \widehat{\lambda})$, and likewise with derivatives such as $\ell_\theta = \partial\ell(\theta)/\partial\theta|_{\theta=\theta_g}$, $\ell_{\lambda\psi} = \partial^2\ell(\theta)/\partial\lambda\partial\psi^{\mathrm{T}}|_{\theta=\theta_g}$. We shall also replace matrices such as $\ell_{\theta\theta}$ by large-sample approximations such as $-\imath_{\theta\theta}$; this can be justified by dividing both sides by n and noting that $-\overline{\ell}_{\theta\theta}(\theta_g) \xrightarrow{P} \imath_1(\theta_g)$.

☐ We shall need to express ℓ_θ, ℓ_λ and $\widehat{\lambda}_\psi - \lambda_g$ in terms of $\widehat{\theta} - \theta_g$. Taylor expansion gives

$$0 = \widehat{\ell}_\theta = \ell_\theta + \ell_{\theta\theta}(\widehat{\theta} - \theta_g) + \cdots = \ell_\theta - \imath_{\theta\theta}(\widehat{\theta} - \theta_g) + \cdots,$$

where $\cdots$ denotes terms of smaller order containing third derivatives of ℓ. The λ component of this equation is

$$0 = \ell_\lambda - \imath_{\lambda\psi}(\widehat{\psi} - \psi_g) - \imath_{\lambda\lambda}(\widehat{\lambda} - \lambda_g) + \cdots.$$

Likewise

$$0 = \tilde{\ell}_\lambda = \ell_\lambda + \ell_{\lambda\lambda}(\widehat{\lambda}_\psi - \lambda_g) + \cdots = \ell_\lambda - \imath_{\lambda\lambda}(\widehat{\lambda}_\psi - \lambda_g) + \cdots.$$

Equating the expressions for ℓ_λ from the last two displays gives

$$\ell_\lambda \doteq \imath_{\lambda\psi}(\widehat{\psi} - \psi_g) + \imath_{\lambda\lambda}(\widehat{\lambda} - \lambda_g) = \imath_{\lambda\lambda}(\widehat{\lambda}_\psi - \lambda_g),$$

so

$$\ell_\theta \doteq \imath_{\theta\theta}(\widehat{\theta} - \theta_g), \quad \ell_\lambda \doteq \imath_{\lambda\lambda}(\widehat{\lambda}_\psi - \lambda_g), \quad \widehat{\lambda}_\psi - \lambda_g \doteq \widehat{\lambda} - \lambda_g + \imath_{\lambda\lambda}^{-1}\imath_{\lambda\psi}(\widehat{\psi} - \psi_g).$$

☐ To obtain the quadratic forms we write

$$\begin{aligned} \ell(\widehat{\theta}) &= \ell(\theta_g) + (\widehat{\theta} - \theta_g)^{\mathrm{T}}\ell_\theta + \tfrac{1}{2}(\widehat{\theta} - \theta_g)^{\mathrm{T}}\ell_{\theta\theta}(\widehat{\theta} - \theta_g) + \cdots \\ &\doteq \ell(\theta_g) + (\widehat{\theta} - \theta_g)^{\mathrm{T}}\imath_{\theta\theta}(\widehat{\theta} - \theta_g) - \tfrac{1}{2}(\widehat{\theta} - \theta_g)^{\mathrm{T}}\imath_{\theta\theta}(\widehat{\theta} - \theta_g), \end{aligned}$$

resulting in

$$\begin{aligned} 2\{\ell(\widehat{\theta}) - \ell(\theta_g)\} &\doteq (\widehat{\theta} - \theta_g)^{\mathrm{T}}\imath_{\theta\theta}(\widehat{\theta} - \theta_g) \\ &= (\widehat{\psi} - \psi_g)^{\mathrm{T}}\imath_{\psi\psi}(\widehat{\psi} - \psi_g) + 2(\widehat{\psi} - \psi_g)^{\mathrm{T}}\imath_{\psi\lambda}(\widehat{\lambda} - \lambda_g) + (\widehat{\lambda} - \lambda_g)^{\mathrm{T}}\imath_{\lambda\lambda}(\widehat{\lambda} - \lambda_g), \end{aligned}$$

and likewise

$$\begin{aligned} 2\{\ell(\widehat{\theta}_\psi) - \ell(\theta_g)\} &\doteq (\widehat{\lambda}_\psi - \lambda_g)^{\mathrm{T}}\imath_{\lambda\lambda}(\widehat{\lambda}_\psi - \lambda_g) \\ &\doteq \left\{(\widehat{\lambda} - \lambda_g) + \imath_{\lambda\lambda}^{-1}\imath_{\lambda\psi}(\widehat{\psi} - \psi_g)\right\}^{\mathrm{T}}\imath_{\lambda\lambda}\left\{(\widehat{\lambda} - \lambda_g) + \imath_{\lambda\lambda}^{-1}\imath_{\lambda\psi}(\widehat{\psi} - \psi_g)\right\} \\ &= (\widehat{\psi} - \psi_g)^{\mathrm{T}}\imath_{\psi\lambda}\imath_{\lambda\lambda}^{-1}\imath_{\lambda\psi}(\widehat{\psi} - \psi_g) + 2(\widehat{\psi} - \psi_g)^{\mathrm{T}}\imath_{\psi\lambda}(\widehat{\lambda} - \lambda_g) + (\widehat{\lambda} - \lambda_g)^{\mathrm{T}}\imath_{\lambda\lambda}(\widehat{\lambda} - \lambda_g). \end{aligned}$$

Subtracting the two quadratic forms gives

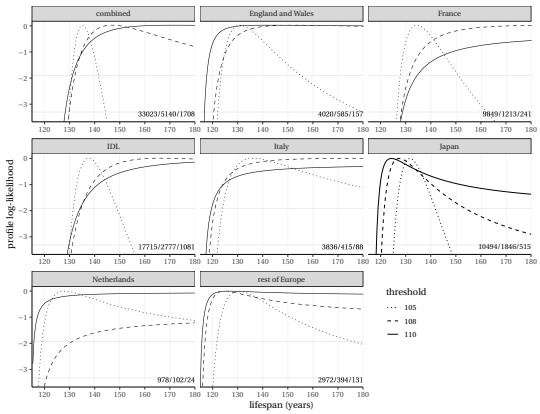
$$\begin{aligned} w_{\mathrm{p}}(\psi_g) &= 2\{\ell(\widehat{\theta}) - \ell(\theta_g)\} - 2\{\ell(\widehat{\theta}_\psi) - \ell(\theta_g)\} \\ &\doteq (\widehat{\psi} - \psi_g)^{\mathrm{T}}(\imath_{\psi\psi} - \imath_{\psi\lambda}\imath_{\lambda\lambda}^{-1}\imath_{\lambda\psi})(\widehat{\psi} - \psi_g), \end{aligned}$$

and as $\widehat{\psi} \mathrel{\dot\sim} \mathcal{N}\{\psi_g, (\imath_{\psi\psi} - \imath_{\psi\lambda}\imath_{\lambda\lambda}^{-1}\imath_{\lambda\psi})^{-1}\}$, we see that $w_{\mathrm{p}}(\psi_g) \mathrel{\dot\sim} \chi^2_p$, as claimed.

☐ Here we are under classical asymptotics, whereby the dimensions of ψ and λ are fixed and $n \to \infty$, and arguments along the lines of Theorem 50 show that the terms $\cdots$ all tend in probability to zero, and thus do not affect the limiting distribution.

Example: Human lifespan

Example 54 *Profile log likelihoods for the endpoint ψ of a generalized Pareto model fitted to data on lifetimes of persons aged over 105 from different databases, with thresholds at 105, 108, 110 years. Here λ is scalar, so $p = q = 1$, and the horizontal line at $-\frac{1}{2}\chi^2_1(0.95) = -1.92$ indicates 95% confidence regions.*



From Belzile et al. (2022, *Annual Review of Statistics and its Application*).

Model selection

☐ The fact that

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

is minimised when $f = g$ suggested 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 will 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 $\dim(\theta)$ (picture ...)

☐ This leads to choosing m to minimise the **Akaike** or the **network** information criteria

$$\mathrm{AIC}_m = 2\left(d_m - \widehat{\ell}_m\right), \quad \mathrm{NIC}_m = 2\left\{\mathrm{tr}(\widehat{\hbar}_m\widehat{\jmath}_m^{-1}) - \widehat{\ell}_m\right\},$$

where the first takes $\mathrm{tr}(\widehat{\hbar}_m\widehat{\jmath}_m^{-1}) \approx d_m = \dim(\theta_m)$.

Note: Derivation of AIC/NIC

☐ As

$$2\mathrm{E}_g\mathrm{E}_g^+\left[-\log\{f(Y^+;\widehat{\theta})/g(Y^+)\}\right] = 2\mathrm{E}_g^+\left\{\log g(Y^+)\right\} - 2\mathrm{E}_g\mathrm{E}_g^+\left\{\log f(Y^+;\widehat{\theta})\right\},$$

we can ignore the first term in the minimisation over f. An unbiased estimator of the second term would be $-2\ell^+(\widehat{\theta})$, where ℓ^+ is the log likelihood based on Y^+ and $\widehat{\theta}$ is based on Y, but the estimator we have available is $-2\ell(\widehat{\theta})$, in which the log likelihood and $\widehat{\theta}$ are both based on Y. Clearly $\ell(\widehat{\theta})$ is upwardly biased, but by how much?

☐ To find out we consider the Taylor expansion

$$\begin{aligned} 2\ell^+(\widehat{\theta}) &= 2\ell^+(\widehat{\theta}^+) + 2(\widehat{\theta}-\widehat{\theta}^+)^{\mathrm{T}}\ell_\theta^+(\widehat{\theta}^+) + (\widehat{\theta}-\widehat{\theta}^+)^{\mathrm{T}}\ell_{\theta\theta}^+(\widehat{\theta}^+)(\widehat{\theta}-\widehat{\theta}^+) + \cdots \\ &= 2\ell^+(\widehat{\theta}^+) - \mathrm{tr}\left\{(\widehat{\theta}-\widehat{\theta}^+)^{\mathrm{T}}\imath_{\theta\theta}(\theta_g)(\widehat{\theta}-\widehat{\theta}^+)\right\} + \cdots \\ &= 2\ell^+(\widehat{\theta}^+) - \mathrm{tr}\left\{(\widehat{\theta}-\widehat{\theta}^+)(\widehat{\theta}-\widehat{\theta}^+)^{\mathrm{T}}\imath_{\theta\theta}(\theta_g)\right\} + \cdots \end{aligned}$$

where $\widehat{\theta}^+$ maximises $\ell^+(\theta)$, $\widehat{\theta}$ maximises $\ell(\theta)$, we have replaced $-\ell_{\theta\theta}^+(\widehat{\theta}^+)$ by its large-sample limit $\imath_{\theta\theta}(\theta_g)$ and neglected terms that are $o_p(1)$. Recall that θ_g is the large-sample limit of $\widehat{\theta}$ when data are sampled from g.

☐ Now $\widehat{\theta}^+$ and $\widehat{\theta}$ are independent and approximately $\mathcal{N}_d(\theta_g, V)$, where $V = \imath_{\theta\theta}^{-1}(\theta_g)\hbar(\theta_g)\imath_{\theta\theta}^{-1}(\theta_g)$, so $\widehat{\theta}^+ - \widehat{\theta} \mathrel{\dot\sim} \mathcal{N}_d(0, 2V)$, giving

$$\begin{aligned} -2\mathrm{E}_g\mathrm{E}_g^+\left\{\ell^+(\widehat{\theta})\right\} &\doteq -2\mathrm{E}_g\mathrm{E}_g^+\left\{\ell(\widehat{\theta})\right\} + \mathrm{tr}\left\{2V\imath_{\theta\theta}(\theta_g)\right\} + o(1) \\ &= 2\left[\mathrm{tr}\left\{\hbar(\theta_g)\imath_{\theta\theta}^{-1}(\theta_g)\right\} - \mathrm{E}_g\mathrm{E}_g^+\left\{\ell(\widehat{\theta})\right\}\right] + o(1). \end{aligned}$$

☐ If $\hbar(\theta_g) \doteq \imath_{\theta\theta}(\theta_g)$, then this final expression can be estimated by $\mathrm{AIC} = 2\{d - \ell(\widehat{\theta})\}$, where $d = \dim(\theta)$, or by the *network information criterion* $\mathrm{NIC} = 2\{\mathrm{tr}(\widehat{\hbar}\widehat{\jmath}^{-1}) - \ell(\widehat{\theta})\}$.

☐ Neither AIC or NIC gives consistent selection of the true model, which would require the penalty to grow with n.

☐ The calculations above use generic large-sample likelihood approximations, and can be improved in specific cases (e.g., with normal errors).

3.3 Nuisance Parameters

Effect of nuisance parameters

Example 55 (Neyman–Scott) *Find the profile log likelihood for σ^2 when $(y_{j1}, y_{j2}) \overset{\mathrm{iid}}{\sim} \mathcal{N}(\mu_j, \sigma^2)$, for $j = 1, \ldots, n$. Comment.*

☐ Profiling over many nuisance parameters can lead to completely wrong inferences, as the previous example shows.

☐ 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.

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

Note to Example 55

☐ The overall log likelihood is

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

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.

☐ In this example there are $d = n+1$ parameters of which n are nuisance parameters.

Dealing with 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 (recall Lemmas 39 and 40) — 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** or a **modified likelihood root**, 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.

☐ We have already seen examples of marginal and conditional likelihoods.

☐ Below we sketch some of the other approaches.

Modified profile likelihood

- 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

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

  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.
- Simpler to base a likelihood on the normal distribution of the modified likelihood root $r^*(\psi)$ (next).

Higher-order inference . . .

- Classical theory gives first-order accuracy, i.e., with ψ scalar

$$\mathrm{P}\left\{r(\psi_g) \leq r^{\mathrm{o}}(\psi_g)\right\} = \Phi\{r^{\mathrm{o}}(\psi)\} + O(n^{-1/2}),$$

  so tests and one-sided confidence sets

$$\{\psi : r^{\mathrm{o}}(\psi) \leq z_{1-\alpha}\}$$

  based on the observed data y^{o} have error $n^{-1/2}$.
- If we replace $r(\psi)$ by the **modified likelihood root**,

$$r^*(\psi) = r(\psi) + \frac{1}{r(\psi)}\log\left\{\frac{q(\psi)}{r(\psi)}\right\},$$

  where $q(\psi)$ depends on the model, then for continuous responses the error drops to $O(n^{-3/2})$, so

$$\mathrm{P}\left\{r^*(\psi_g) \leq r^{*\mathrm{o}}(\psi_g)\right\} = \Phi\{r^{*\mathrm{o}}(\psi_g)\} + O(n^{-3/2}),$$

  so a one-sided confidence set

$$\{\psi : r^{*\mathrm{o}}(\psi) \leq z_{1-\alpha}\}$$

  has error of order $n^{-3/2}$; often this almost exact even for tiny n (recall Example 52).

. . . with nuisance parameters

- With nuisance parameters, $r(\psi) = \text{sign}(\widehat{\psi} - \psi)\sqrt{w_{\text{p}}(\psi)}$, and

$$q(\psi) = \frac{|\, \varphi(\widehat{\theta}) - \varphi(\widehat{\theta}_\psi) \;\; \varphi_\lambda(\widehat{\theta}_\psi) \,|}{|\, \varphi_\theta(\widehat{\theta}) \,|} \left\{ \frac{|\, \widehat{\jmath} \,|}{|\, \jmath_{\lambda\lambda}(\widehat{\theta}_\psi) \,|} \right\}^{1/2}$$

  where φ is the $d \times 1$ canonical parameter of a local **exponential family approximation** to the model at the observed data y^{o}, with $\varphi_\theta(\theta) = \partial\varphi(\theta)/\partial\theta^{\text{T}}$, etc.

- In a general exponential family $\varphi(\theta)$ is the canonical parameter, and in a linear exponential family,

$$q(\psi) = (\widehat{\psi} - \psi) \left\{ \frac{|\widehat{\jmath}|}{|\jmath_{\lambda\lambda}(\widehat{\theta}_\psi)|} \right\}^{1/2}.$$

- In general for independent continuous observations we write

$$\varphi(\theta)_{d\times 1} = V^{\text{T}}_{d\times n} \left. \frac{\partial \ell(\theta; y)}{\partial y} \right|_{y=y^{\text{o}}} = \sum_{j=1}^{n} V_j^{\text{T}} \left. \frac{\partial \log f(y_j; \psi, \lambda)}{\partial y_j} \right|_{y=y^{\text{o}}},$$

  where the $1 \times d$ vectors $V_j = \partial y_j/\partial\theta^{\text{T}}$ are evaluated at y^{o} and $\widehat{\theta}^{\text{o}}$.

Properties of higher order approximations

- Invariant to interest-respecting reparameterization.
- Computation almost as easy as first order versions.
- Error $O(n^{-3/2})$ in continuous response models, $O(n^{-1})$ in discrete response models.
- Relative (not absolute) error, so highly accurate in tails.
- Bayesian version is also available (and easier to derive).

Example 56 (Location-scale model) *Compute $\varphi(\theta)$ for a location-scale model, in which independent observations Y_j have density $\tau^{-1}h\{(y-\eta)/\tau\}$. What about the normal density?*

Note to Example 56

- In this case the overall log likelihood is

$$\ell(\eta, \tau) = -n \log \tau + \sum_{j=1}^{n} \log h\{(y_j - \eta)/\tau\},$$

so the vector $\partial \ell(\eta, \tau)/\partial y$ has components $\tau^{-1}(\log h)'\{(y_j - \eta)/\tau\}$, evaluated at the parameters η and τ and observed data vector $y_1^{\mathrm{o}}, \ldots, y_n^{\mathrm{o}}$.

- To compute the V_j we use the structural expression $y = \eta + \tau\varepsilon$, where $\varepsilon \sim h$. This represents y as a function of $\theta^{\mathrm{T}} = (\eta, \tau)$, and yields $\partial y_j/\partial\theta^{\mathrm{T}} = (1, \varepsilon_j)$. This has to be evaluated at the observed data point y^{o}, and at that point the parameters are replaced by their maximum likelihood estimates, giving $V_j^{\mathrm{T}} = (1, (y_j^{\mathrm{o}} - \widehat{\eta}^{\mathrm{o}})/\widehat{\tau}^{\mathrm{o}})$.

- This yields

$$\varphi(\theta) = \sum_{j=1}^{n} \tau^{-1}(\log h)'\{(y_j^{\mathrm{o}} - \eta)/\tau\}(1, \varepsilon_j^{\mathrm{o}})^{\mathrm{T}},$$

where we have set $\varepsilon_j^{\mathrm{o}} = (y_j^{\mathrm{o}} - \widehat{\eta}^{\mathrm{o}})/\widehat{\tau}^{\mathrm{o}}$.

- If h is normal, then $\log h(u) \equiv -u^2/2$, so $(\log h)'\{(y_j^{\mathrm{o}} - \eta)/\tau\} = -(y_j^{\mathrm{o}} - \eta)/\tau^2$, leading to

$$\varphi(\theta)^{\mathrm{T}} = \left(\sum_{j=1}^{n} (\eta - y_j^{\mathrm{o}})/\tau^2, \sum_{j=1}^{n} (\eta - y_j^{\mathrm{o}})/\tau^2 \times e_j \right) \equiv (\eta/\tau^2, 1/\tau^2),$$

because it turns out that inferences are invariant under non-singular affine transformations of $\varphi(\theta)$ (exercise).

Orthogonal parameters

- If the expected information matrix is block diagonal, with $\imath_{\psi,\lambda}(\theta) = 0$ for all θ, then $\widehat{\psi}$ is asymptotically independent of $\widehat{\lambda}$, and 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 $\widehat{\jmath}_{\lambda\psi} \approx 0$, for example by parametrising it so that $\imath_{\lambda\psi}(\theta) \equiv 0$, 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.

Orthogonalisation

☐ 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.

Orthogonal 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.

Example 57 (Linear exponential family) *What parameter is orthogonal to ψ in the linear exponential family with log likelihood*

$$\ell^*(\psi, \gamma) \equiv s_1^{\mathrm{T}} \psi + s_2^{\mathrm{T}} \gamma - k(\psi, \gamma)?$$

Consider normal and Poisson likelihoods in particular.

Note to Example 57

☐ The parameters $\lambda = \lambda(\psi, \gamma)$ orthogonal to ψ are determined by

$$\frac{\partial \gamma}{\partial \psi^{\mathrm{T}}} = -k_{\gamma\gamma}^{-1}(\psi, \gamma) k_{\gamma\psi}(\psi, \gamma). \tag{4}$$

If we reparametrize in terms of ψ and $\lambda = k_\gamma(\psi, \gamma) = \partial k(\psi, \gamma)/\partial\gamma$, then in this new parametrization, γ is a function of ψ and λ, and

$$0 = \frac{\partial \lambda^{\mathrm{T}}}{\partial \psi} = \frac{\partial \gamma^{\mathrm{T}}}{\partial \psi} k_{\gamma\gamma}(\psi, \gamma) + k_{\psi\gamma}(\psi, \gamma),$$

so $\lambda = k_\gamma(\psi, \gamma)$ is a solution to (4). That is, the parameter orthogonal to ψ is the so-called complementary mean parameter $\lambda(\psi, \gamma) = \mathrm{E}(S_2; \psi, \gamma)$. By symmetry, $\mathrm{E}(S_1; \psi, \gamma)$ is orthogonal to γ.

☐ The normal distribution with mean μ and variance σ^2 has canonical parameter $(\mu/\sigma^2, -1/(2\sigma^2))$. The canonical statistic (Y, Y^2) has expectation $(\mu, \mu^2 + \sigma^2)$, so μ is orthogonal to $-1/(2\sigma^2)$, and hence to σ^2, while μ/σ^2 is orthogonal to $\mu^2 + \sigma^2$.

☐ Independent Poisson variables Y_1 and Y_2 with means $\exp(\gamma)$ and $\exp(\gamma + \psi)$ have log likelihood

$$\ell^*(\psi, \gamma) \equiv (y_1 + y_2)\gamma + y_2\psi - e^\gamma - e^{\gamma+\psi}.$$

The discussion above suggests that

$$\lambda = \mathrm{E}(Y_1 + Y_2) = \exp(\gamma) + \exp(\gamma + \psi) = e^\gamma(1 + e^\psi)$$

is orthogonal to ψ, so $\gamma = \log\lambda - \log(1 + e^\psi)$ and

$$\ell(\psi, \lambda) \equiv y_2\psi - (y_1 + y_2)\log(1 + e^\psi) + (y_1 + y_2)\log\lambda - \lambda.$$

The separation of ψ and λ implies that the profile and modified profile likelihoods for ψ are proportional. They correspond to the conditional likelihood obtained from the density of Y_2 given $Y_1 + Y_2$.

Composite likelihood

- 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 $\tilde{\theta}$, but requires a sandwich variance matrix (or some other approach) to estimate $\mathrm{var}(\tilde{\theta})$.
- Model comparisons use the **composite likelihood information criterion**

$$\mathrm{CLIC} = 2\left[\mathrm{tr}\{\hbar(\tilde{\theta})\jmath(\tilde{\theta})^{-1}\} - \ell_{\mathrm{C}}(\tilde{\theta})\right].$$

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
  - **empirical likelihood**, an extension of nonparametric modelling (econometrics).
- 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.