Topics in Likelihood Inference

STA4508H

Nancy Reid
University of Toronto

January 19, 2022


CITY OF TORONTO WORKPLACE BY-LAW
MAXIMUM PENALTY $2,000.00
STATISTIC
CINTAS
416.798.9000

Various 'types' of likelihood

1. likelihood, marginal and conditional likelihood, profile likelihood, adjusted profile
2. semi-parametric likelihood, partial likelihood
3. quasi-likelihood, composite likelihood — misspecified models
4. empirical likelihood, penalized likelihood
5. simulated likelihood, indirect inference
6. bootstrap likelihood, h-likelihood, weighted likelihood, pseudo-likelihood, local likelihood, sieve likelihood

Recap

- likelihood function is proportional to the probability of the observed data
- need to assume a probability model in order to write down a likelihood function
- these models are usually parametric, i.e. a class of models that vary with a parameter $\theta \in \Theta$
- but are sometimes non-parametric, in the sense that Θ might be an infinite-dimensional space

e.g. the class of all twice-differentiable function

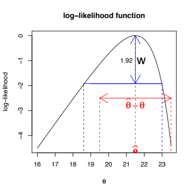
e.g. the intensity function for a Poisson process

- random effects model: why do we integrate out the random effects?

- several examples: regression, time series, continuous time processes, correlated binary data, etc.
- several examples of likelihood functions that involve integration — complicated
- an example where the likelihood function can't be written down completely — Ising model

- these examples meant to motivate variations on the usual likelihood function — to come

- notation and derived quantities: score function, observed and expected Fisher information, maximum likelihood estimate, likelihood ratio statistic



Don't forget the handout

STA 4508: Likelihood and derived quantities January 2022

Given a model for Y which assumes Y has a density $f(y;\theta), \quad \theta \in \Theta \subset \mathbb{R}^d$, we have the following definitions:

observed likelihood function	$L(\theta;y) = c(y)f(y;\theta)$
log-likelihood function	$\ell(\theta;y) = \log L(\theta;y) = \log f(y;\theta) + a(y)$
score function	$U(\theta) = \partial\ell(\theta;y)/\partial\theta$
observed information function	$j(\theta) = -\partial^2\ell(\theta;y)/\partial\theta\partial\theta^T$
expected information (in one observation)	$i(\theta) = \mathrm{E}_\theta U(\theta)U(\theta)^T$ (called $i_1(\theta)$ in CH)

When we have Y_i independent, identically distributed from $f(y_i;\theta)$, then, denoting the observed sample $y = (y_1, \ldots, y_n)$ we have:

log-likelihood function	$\ell(\theta) = \ell(\theta;y) + a(y)$	$O_p(n)$
maximum likelihood estimate	$\hat\theta = \hat\theta(y) = \arg\sup_\theta \ell(\theta)$	$\theta + O_p(n^{-1/2})$
score function	$U(\theta) = \ell'(\theta) = \sum U_i(\theta) = U_+(\theta)$	$O_p(n^{1/2})$
observed information function	$j(\theta) = -\ell''(\theta) = -\ell(\theta;Y)$	$O_p(n)$
observed (Fisher) information	$j(\hat\theta)$	
expected (Fisher) information	$i(\theta) = E_\theta\{U(\theta)U(\theta)^T\} = ni_1(\theta)$	$O(n)$,

where with the risk of some confusion we use the same notation. Sometimes the expected Fisher information is defined instead as $i(\theta) = \mathrm{E}_\theta\{-\partial U(\theta;Y)/\partial\theta^T\}$ (e.g.

... Recap: inference based on likelihood

- "pure likelihood": values of θ are plausible if $L(\hat{\theta})/L(\theta)$ not too large

  or $L(\theta)/L(\hat{\theta})$ not too small

- Bayesian inference: posterior $\propto$ Likelihood $\times$ prior $\qquad \pi(\theta \mid y) \propto L(\theta; y)\pi(\theta)$

- frequentist: quantities derived from the likelihood function have "good" properties

  behave well when we have large samples from the model

- also frequentist: pivotal quantities derived from the likelihood function can be used to construct p-value functions

  also called significance functions
- p-value functions provide nested sets of confidence intervals

  if monotone in θ

A trio of limit results

1.

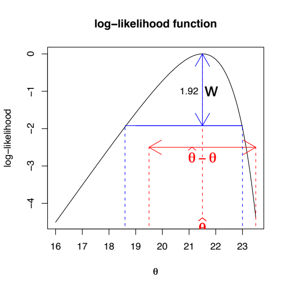
$$\frac{1}{\sqrt{n}} U(\theta) \xrightarrow{d} N\{0, i_1(\theta)\}$$

2.

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\{0, i_1^{-1}(\theta)\}$$

3.

$$2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{d} \chi_1^2$$



Leading to a trio of approximate confidence intervals:

1. $\{\theta : |U(\theta) i^{-1/2}(\theta)| \leq z_{1-\alpha/2}\}$

2. $\{\theta : |(\hat{\theta} - \theta) i^{1/2}(\hat{\theta})| \leq z_{1-\alpha/2}\}$

3. $\{\theta : 2[\ell(\hat{\theta}) - \ell(\theta)]\} \leq \chi^2_{1,1-\alpha}$

- or leading to a trio of approximate pivotal quantities

$$
\begin{aligned}
1.\quad r_u(\theta) &= U(\theta) j^{-1/2}(\hat{\theta}) \overset{.}{\sim} N(0,1) \\
2.\quad r_e(\theta) &= (\hat{\theta} - \theta) j^{1/2}(\hat{\theta}), \\
3.\quad r(\theta) &= \text{sign}(\hat{\theta} - \theta)[2\{\ell(\hat{\theta}) - \ell(\theta)\}]^{1/2}
\end{aligned}
$$

- $\Pr\{r_u(\theta) \le r_u^0(\theta)\} \doteq \Phi\{r_u^0(\theta)\}$ under sampling from the model $f(y;\theta) = f(y_1, \dots, y_n; \theta)$

- and a trio of p-value functions of θ, for fixed data
- similarly

$$
\begin{aligned}
1.\quad p_u(\theta) &= \Phi\{r_u^0(\theta)\}, \\
2.\quad p_e(\theta) &= \Phi\{r_e(\theta)\} \\
3.\quad p_r(\theta) &= \Phi\{r(\theta)\}
\end{aligned}
$$

Observed and expected Fisher information

Example: Exponential

- $f(y_i;\theta) = \theta e^{-y_i\theta}, \quad i = 1,\dots,n$
- $\ell(\theta) =$
- $\ell'(\theta) =$
- $\ell''(\theta) =$
- $r_u(\theta) =$
- $r_e(\theta) =$
- $r(\theta) =$

expand $\log(\theta\bar{y})$ around 1 to get asymptotic equivalence to r_e, r_u

Example: Exponential

- $f(y_i;\theta) = \theta e^{-y_i\theta}, \quad i = 1, \dots, n$
- $\ell(\theta) = n\log\theta - n\theta\bar{y}$
- $\ell'(\theta) = \frac{n}{\theta} - n\bar{y}$ $\qquad\hat{\theta} = \bar{y}^{-1}$
- $\ell''(\theta) = -\frac{n}{\theta^2}$
- $r_u(\theta) = \frac{1}{\sqrt{n}}\ell'(\theta)j^{-1/2}(\hat{\theta}) = \sqrt{n}(\frac{1}{\theta\bar{y}} - 1)$
- $r_e(\theta) = (\hat{\theta} - \theta)j^{1/2}(\hat{\theta}) = \sqrt{n}(1 - \bar{y}\theta)$
- $r(\theta) = \sqrt{(2n)}\{\theta\bar{y} - 1 - \log(\theta\bar{y})\}^{1/2}$

expand $\log(\theta\bar{y})$ around 1 to get asymptotic equivalence to r_e, r_u

... Example: Exponential

14 CHAPTER 2. UNCERTAINTY AND APPROXIMATION

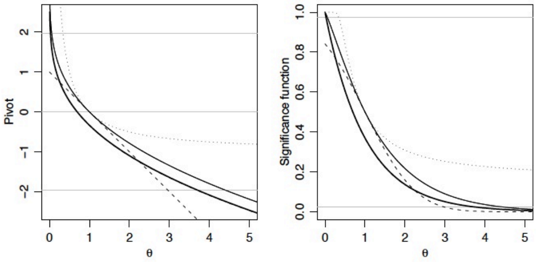


Figure 2.2: Approximate pivots and P-values based on an exponential sample of size $n = 1$. Left: likelihood root $r(\theta)$ (solid), score pivot $s(\theta)$ (dots), Wald pivot $t(\theta)$ (dashes), modified likelihood root $r^*(\theta)$ (heavy), and exact pivot $\theta \sum y_j$ (dot-dash). The modified likelihood root is indistinguishable from the exact pivot. The horizontal lines are at $0, \pm 1.96$. Right: corresponding significance functions, with horizontal lines at 0.025 and 0.975.

Aside

- for inference re θ, given y, plot $p(\theta)$ vs θ
- for p-value for $H_0 : \theta = \theta_0$, compute $p(\theta_0)$
- for checking whether, e.g. $\Phi\{r_e(\theta)\}$ is a good approximation,
  - compare $p(\theta) = \Phi\{r_e(\theta)\}$ to $p_{\text{exact}}(\theta)$, as a function of θ, fixed y
  - or compare $p(\theta_0)$ to $p_{\text{exact}}(\theta_0)$ as a function of y
- if $p_{\text{exact}}(\theta)$ not available, simulate
- if θ is a vector, choose one component at a time

Vector parameter limit theorems and approximations

- $U(\theta)$

- $\hat{\theta}$

- $2\{\ell(\hat{\theta}) - \ell(\theta)\}$

Parameter of interest and nuisance parameter

- $\theta = (\psi, \lambda) =$
- $U(\theta) =$
- $i(\theta) =$ $j(\theta) =$
- $i^{-1}(\theta) =$ $j^{-1}(\theta) =$
- $i^{\psi\psi}(\theta) =$
- $\ell_{\mathrm{P}}(\psi) =$ $j_{\mathrm{P}}(\psi) =$

Nuisance parameters

- $\theta = (\psi, \lambda) = (\psi_1, \ldots, \psi_q, \lambda_1, \ldots, \lambda_{d-q})$

- $U(\theta) = \begin{pmatrix} U_\psi(\theta) \\ U_\lambda(\theta) \end{pmatrix}, \qquad U_\lambda(\psi, \hat{\lambda}_\psi) = 0$
- $i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix} \quad j(\theta) = \begin{pmatrix} j_{\psi\psi} & j_{\psi\lambda} \\ j_{\lambda\psi} & j_{\lambda\lambda} \end{pmatrix}$
- $i^{-1}(\theta) = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix} \quad j^{-1}(\theta) = \begin{pmatrix} j^{\psi\psi} & j^{\psi\lambda} \\ j^{\lambda\psi} & j^{\lambda\lambda} \end{pmatrix}.$
- $i^{\psi\psi}(\theta) = \{i_{\psi\psi}(\theta) - i_{\psi\lambda}(\theta) i^{-1}_{\lambda\lambda}(\theta) i_{\lambda\psi}(\theta)\}^{-1},$

- $\ell_{\mathrm{P}}(\psi) = \ell(\psi, \hat{\lambda}_\psi), \qquad j_{\mathrm{P}}(\psi) = -\ell''_{\mathrm{P}}(\psi)$

Approximations from limiting distributions, nuisance parameters

$$
\begin{aligned}
w_u(\psi) = U_\psi(\psi, \hat\lambda_\psi)^T \{i^{\psi\psi}(\psi, \hat\lambda_\psi)\} U_\psi(\psi, \hat\lambda_\psi) \quad &\dot\sim \quad \chi^2_q \\
w_e(\psi) = (\hat\psi - \psi)\{i^{\psi\psi}(\hat\psi, \hat\lambda)\}^{-1}(\hat\psi - \psi) \quad &\dot\sim \quad \chi^2_q \\
w(\psi) = 2\{\ell(\hat\psi, \hat\lambda) - \ell(\psi, \hat\lambda_\psi)\} = 2\{\ell_\mathrm{P}(\hat\psi) - \ell_\mathrm{P}(\psi)\} \quad &\dot\sim \quad \chi^2_q;
\end{aligned}
$$

Approximate Pivots, $q = 1$

$$
\begin{aligned}
r_u(\psi) &= \ell'_\mathrm{P}(\psi) j_\mathrm{P}(\hat\psi)^{-1/2} \mathrel{\dot\sim} N(0, 1), \\
r_e(\psi) &= (\hat\psi - \psi) j_\mathrm{P}(\hat\psi)^{1/2} \mathrel{\dot\sim} N(0, 1), \\
r(\psi) &= \mathrm{sign}(\hat\psi - \psi)[2\{\ell_\mathrm{P}(\hat\psi) - \ell_\mathrm{P}(\psi)\}]^{1/2} \mathrel{\dot\sim} N(0, 1)
\end{aligned}
$$

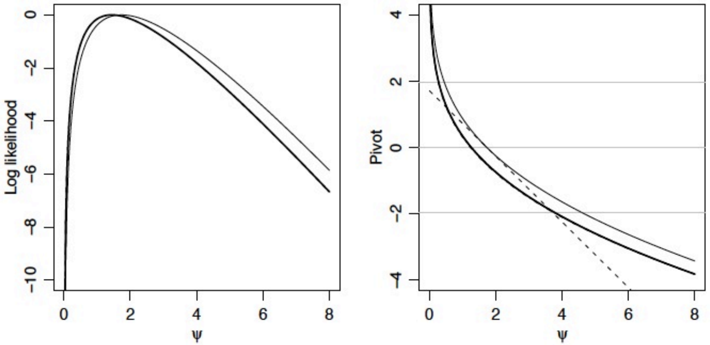


Figure 2.3: Inference for shape parameter ψ of gamma sample of size $n = 5$. Left: profile log likelihood $\ell_{\rm p}$ (solid) and the log likelihood from the conditional density of u given v (heavy). Right: likelihood root $r(\psi)$ (solid), Wald pivot $t(\psi)$ (dashes), modified likelihood root $r^*(\psi)$ (heavy), and exact pivot overlying $r^*(\psi)$. The horizontal lines are at $0, \pm 1.96$.

Properties of likelihood functions and likelihood inference

- the likelihood depends only on the minimal sufficient statistic
- recall: $L(\theta; y) = m_1(s; \theta)m_2(y) \iff s$ is minimal sufficient
- equivalently $\dfrac{L(\theta; y)}{L(\theta_0; y)}$ depends only on s
- " the likelihood map is sufficient" Fraser & Naderi, 2006; Barndorff-Nielsen, et al, 1976

  i.e $y \to \bar{L}_0(\cdot; y)$, or $y \to \bar{L}(\cdot; y)$ normed

... properties

- maximum likelihood estimates are equivariant: $\hat{h}(\theta) = h(\hat{\theta})$ for one-to-one $h(\cdot)$

- question: which of r_e, r_u, r are invariant under interest-respecting reparameterizations $(\psi, \lambda) \to \{\psi, \eta(\psi, \lambda)\}$?

- consistency of maximum likelihood estimate?
- equivalence of maximum likelihood estimate and root of score equation?

- observed vs. expected information

Approximate Bayesian inference

- $\pi(\theta \mid y) = \dfrac{\exp\{\ell(\theta; y)\}\pi(\theta)}{\int \exp\{\ell(\theta; y)\}\pi(\theta)d\theta}$

- expand numerator and denominator about $\hat{\theta}$, assuming $\ell'(\hat{\theta}) = 0$

- $\pi(\theta \mid y) \doteq N\{\hat{\theta}, j^{-1}(\hat{\theta})\}$

Profile likelihood: examples

- regression

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \quad \psi = \sigma^2$$

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^T(y - X\hat{\beta})$$

- Neyman-Scott

$$y_{ij} \sim N(\mu_i, \sigma^2), j = 1, \ldots, k; i = 1, \ldots, m$$

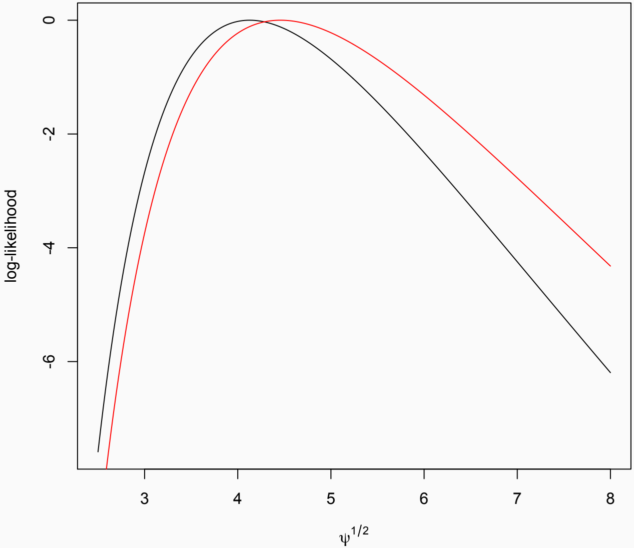
$$\hat{\sigma}^2 = \frac{1}{mk}\sum_{i=1}^{m}(y_{ij} - \bar{y}_{i.})^2$$

- 2×2 tables

$$y_{i1} \sim Bern(p_{i1}), y_{i2} \sim Bern(p_{i2}), i = 1, \ldots, n, \quad \log\{\frac{p_{i1}/(1 - p_{i1})}{p_{i2}(1 - p_{i2})}\} = \psi + \lambda_i$$

$$\hat{\psi} \xrightarrow{p} \psi/2$$

This is a plot of $-n \log \sigma - (y - X\hat{\beta})^T(y - X\hat{\beta})/2\sigma^2$ (black), and $-(n-p) \log \sigma - (y - X\hat{\beta})^T(y - X\hat{\beta})/2\sigma^2$ against σ (red) for given data



Eliminating nuisance parameters

- Profile likelihood poor if q large; fails if $q \to \infty$

- alternative: marginal likelihood: $f(\underline{y}_n; \psi, \lambda) \propto f_{\mathrm{m}}(t_1; \psi) f_{\mathrm{c}}(t_2 \mid t_1; \psi, \lambda)$ $\qquad t_j = t_j(\underline{y})$

- Example $N(X\beta, \sigma^2 I)$: $\quad f(y; \beta, \sigma^2) \propto f_{\mathrm{m}}(RSS; \sigma^2) f_{\mathrm{c}}(\hat{\beta} \mid RSS; \beta, \sigma^2)$

$$L_m(\sigma^2) \propto f_m(RSS; \sigma^2)$$

- alternative conditional likelihood: $f(\underline{y}; \psi, \lambda) \propto f_{\mathrm{c}}(t_1 \mid t_2; \psi) f_{\mathrm{m}}(t_2; \psi, \lambda)$

- Example 2×2 tables: $f(\underline{y}; \psi, \underline{\lambda}) \propto \prod_{i=1}^n f_c(y_{i1} \mid y_{i1} + y_{i2}; \psi) f_m(y_{i1} + y_{i2}; \psi, \lambda_i)$

$$L_c(\psi) = \prod f_c(y_{i1} \mid y_{i1} + y_{i2}; \psi)$$

Linear exponential families

- conditional density free of nuisance parameter
- $f(y_i;\psi,\lambda) = \exp\{\psi^T s(y_i) + \lambda^T t(y_i) - k(\psi,\lambda)\} h(y_i)$
- $f(y;\psi,\lambda) = \exp\{\psi^T \Sigma s(y_i) + \lambda^T \Sigma t(y_i) - nk(\psi,\lambda)\} \Pi h(y_i)$

  Let $s = \Sigma s(y_i), t = \Sigma t(y_i)$
- $f(s,t;\psi,\lambda) = \exp\{\psi^T s + \lambda^T t - nk(\psi,\lambda)\}\tilde{h}(s)$

$$
\begin{aligned}
f(s \mid t;\psi) &= \frac{f(s,t;\psi,\lambda)}{\int f(s,t;\psi,\lambda) ds} \\
&= \frac{\exp\{\psi^T s + \lambda^T t - nk(\psi,\lambda)\}\tilde{h}(s)}{\int \exp\{\psi^T s + \lambda^T t - nk(\psi,\lambda)\}\tilde{h}(s) ds} \\
&= \frac{\exp\{\psi^T s\}\tilde{h}(s)}{\int \exp\{\psi^T s\}\tilde{h}(s) ds} \\
&= \exp\{\psi^T s - n\tilde{k}_t(\psi)\}\tilde{h}_t(s)
\end{aligned}
$$

$\tilde{k}_t$, $\tilde{h}_t$ just convenient notation for integral of denominator

Logistic regression

- $y_i \sim Binom(m_i, p_i), i = 1, \ldots, n$
- $\log\{p_i/(1 - p_i)\} = x_i^{\mathrm{T}}\beta$
- $f(y;\beta) = \exp\{\beta_1\Sigma(x_{i1}y_i) + \cdots + \beta_p\Sigma(x_{ip}y_i) - \Sigma m_i \log(1 + e^{x_i^{\mathrm{T}}\beta})\}$
- $f_c(s_5 \mid s_{-(5)}; \beta_5) \propto \exp\{\beta_5 s_5 - \tilde{k}(\beta_5)\}h(s)$

… Logistic regression

4.2. *URINE DATA* 57

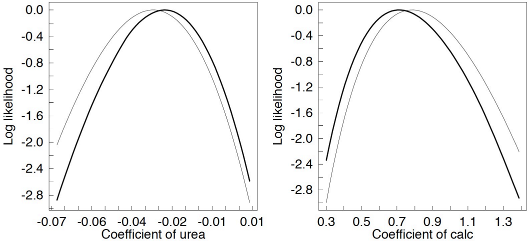


Figure 4.2: Comparison of log likelihoods for the urine data: profile log likelihood (solid line), approximate conditional log likelihood (bold line). The variables of interest are urea (left panel) and calcium concentration (right panel). The graphical output is obtained with the `plot` method of the `cond` package.

$$f_c(s_5 \mid s_{-(5)}; \beta_5) \propto \exp\{\beta_5 s_5 - \tilde{k}(\beta_5)\} h(s)$$

Summary 4.1 Approximate conditional inference for the urine data.

```
> urine.glm <- glm( formula=r~I(100*(gravity-1))+ph+osmo+conduct+urea+calc,
+                   family=binomial, data=urine )

> summary(urine.glm)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             0.60609    3.79582   0.160  0.87314
I(100 * (gravity - 1))  3.55944    2.22110   1.603  0.10903
ph                     -0.49570    0.56976  -0.870  0.38429
osmo                    0.01681    0.01782   0.944  0.34536
conduct                -0.43282    0.25123  -1.723  0.08493 .
urea                   -0.03201    0.01612  -1.986  0.04703 *
calc                    0.78369    0.24216   3.236  0.00121 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

    Null deviance: 105.17  on 76  degrees of freedom
Residual deviance:  57.56  on 70  degrees of freedom
AIC: 71.56

> urine.cond.urea <- cond( urine.glm, offset=urea )

> coef( urine.cond.urea )
           Estimate   Std. Error
uncond.  -0.03201315   0.01611884
cond.    -0.02759202   0.01489919

> summary( urine.cond.urea, coef=F )

Confidence intervals
--------------------
 level = 95 %
                                           lower two-sided      upper
Wald pivot                                -0.06361          -0.0004208
```

Summary 4.1 Approximate conditional inference for the urine data (cont.).

```
> urine.cond.calc <- cond( urine.glm, offset=calc )

> coef( urine.cond.calc )
          Estimate  Std. Error
uncond.   0.7836913   0.2421638
cond.     0.7110584   0.2282501

> summary( urine.cond.calc, coef=F )

Confidence intervals
 level = 95 %
                                         lower two-sided upper
Wald pivot                              0.3091           1.258
Wald pivot (cond. MLE)                  0.2637           1.158
Likelihood root                         0.3815           1.342
Modified likelihood root                0.3193           1.213
Modified likelihood root (cont. corr.)  0.3044           1.254

Diagnostics:
-----------
    INF      NP
0.08451 0.32878
```

Marginal and conditional likelihoods

$$
\begin{aligned}
L_c(\psi) &= \log f_c\{s(y) \mid t(y); \psi\}, \\
L_m(\psi) &= \log f_m\{s(y); \psi\}
\end{aligned}
$$

- Inference based on usual asymptotics applies, under regularity conditions on $f(y; \psi, \lambda)$
- likelihoods based on observable random variables
- Bartlett identities apply directly
- use conditional or marginal Fisher information, etc.

- might lose information in other component

$$f(y; \psi, \lambda) \propto f_m(s; \psi) f_c(t \mid s; \psi, \lambda)$$

- marginal likelihoods associated with transformation models

REML

Approximate conditional inference

- $\ell_c(\psi) \doteq \ell_p(\psi) - \frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)|$ $\qquad i_{\psi\lambda}(\theta) = 0$
- $\ell_m(\psi) \doteq \ell_p(\psi) - \frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)|$
- $\ell_c(\psi) \doteq \ell_p(\psi) + \frac{1}{2}\log|j_{\eta\eta}(\psi, \hat\eta_\psi)|$ $\qquad \exp\{\psi^{\mathrm{T}}s + \eta^{\mathrm{T}}t - c(\psi, \eta)\}$
- adjusted profile log-likelihood

$$\ell_A(\psi) = \ell_p(\psi) + A(\psi)$$

$A(\psi)$ assumed to be $O_p(1)$

- generic form is $A_{FR}(\psi) = +\frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)| - \log|\frac{d(\lambda)}{d\hat\lambda_\psi}|$ Fraser 03
- closely related $A_{BN}(\psi) = -\frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)| + \log|\frac{d\hat\lambda}{d\hat\lambda_\psi}|$ SM §12.4.1

BN 83