Likelihood inference in complex settings

Nancy Reid

with Uyen Hoang, Wei Lin, Ximing Xu

FOUNDATIONS AND FRONTIERS:
A Conference Celebrating the Contributions of
Mary Thompson to the Statistical Sciences
October 28 & 29, 2011
University of Waterloo

UNIVERSITY of TORONTO

Why likelihood?

- likelihood function depends on data only through sufficient statistics
- "likelihood map is sufficient" Fraser & Naderi, 2006
- provides summary statistics with known limiting distribution
- leading to approximate pivotal functions, based on normal distribution
- in some models the likelihood function gives exact inference
- "likelihood function as pivotal" Hinkley, 1980
- likelihood function $+$ sample space derivative gives better approximate inference

Summary statistics and approximate pivotals

- model — $f(y;\theta), y \in \mathbb{R}^n, \theta \in \mathbb{R}^d$
- log-likelihood function — $\ell(\theta; y) = \log f(y;\theta) + a(y)$
- score function — $u(\theta) = \partial \ell(\theta; y)/\partial\theta$
- maximum likelihood estimate — $\hat\theta = \arg\sup_\theta \ell(\theta; y)$
- log-likelihood ratio — $w(\theta) = 2\{\ell(\hat\theta; y) - \ell(\theta; y)\}$

Approximate pivotals

$$\sqrt{n}(\hat{\theta} - \theta) \dot{\sim} N_d\{0, j^{-1}(\hat{\theta})\}$$

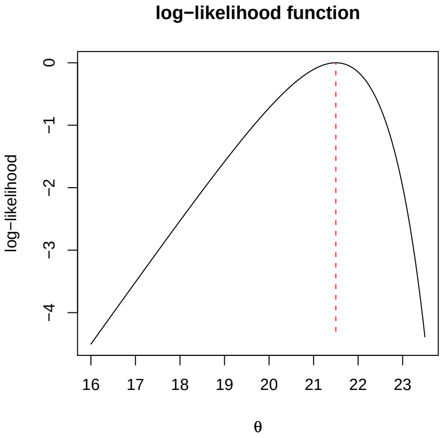
$$w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \dot{\sim} \chi^2_d$$

$$\frac{1}{\sqrt{n}} U(\theta) \dot{\sim} N_d\{0, j(\hat{\theta})\}$$

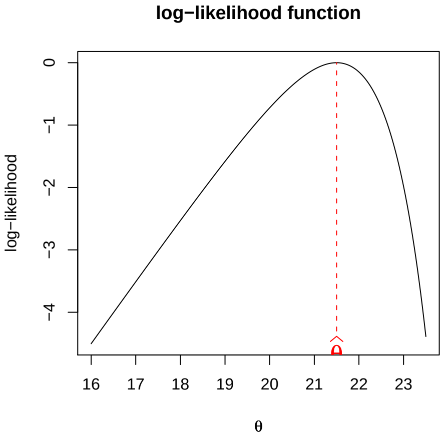
$$\frac{1}{\sqrt{n}} U(\theta) \xrightarrow{\mathcal{L}} N_d\{0, \mathcal{I}(\theta)\}$$

$$j(\hat{\theta}) = -\ell''(\hat{\theta})/n \qquad \mathcal{I}(\theta) = E\{j(\theta)\}$$

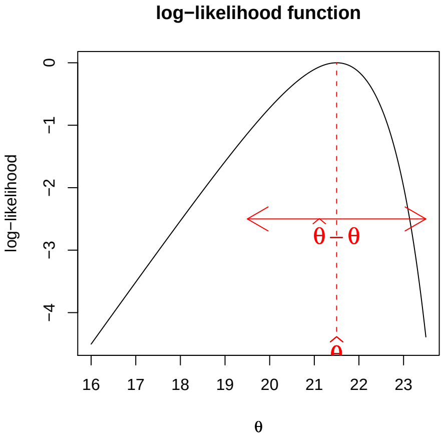
...approximate pivotals



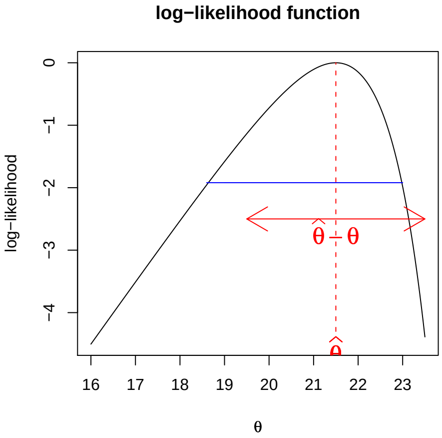
...approximate pivotals



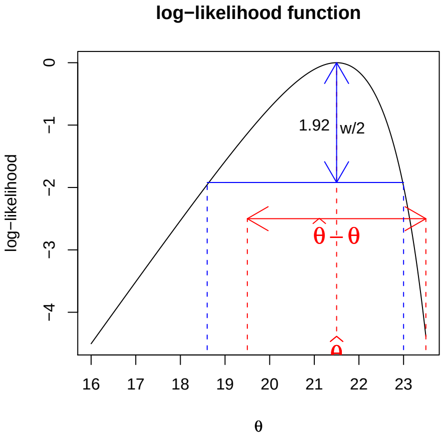
...approximate pivotals



...approximate pivotals



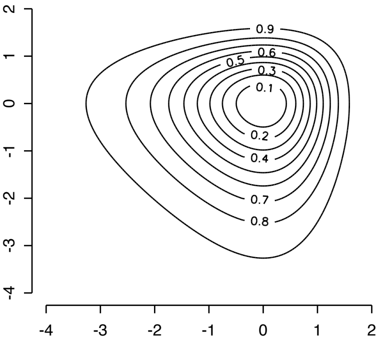
...approximate pivotals



...approximate pivotals

$$w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \dot{\sim} \chi^2_d$$

(a)



Likelihood as pivotal

- Example: location model $f(y;\theta) = \prod_{i=1}^n f_0(y_i - \theta), \quad \theta \in \mathbb{R}$

- Fisher (1934) $\quad f(\hat\theta \mid a;\theta) = \dfrac{\exp\{\ell(\theta;y)\}}{\int \exp\{\ell(\theta;y)\}d\theta}$

- $$(y_1,\ldots,y_n) \longleftrightarrow (\hat\theta, a_1,\ldots,a_n) \qquad a_i = y_i - \hat\theta$$

- exact (conditional) distribution of maximum likelihood estimator given by renormalized likelihood function

- p^* approximation:

$$p^*(\hat\theta \mid a;\theta) = c(\theta,a)|j(\hat\theta)|^{1/2}\exp\{\ell(\theta;\hat\theta,a) - \ell(\hat\theta;\hat\theta,a)\}$$

A simpler approach

- avoid

$$(y_1, \ldots, y_n) \longleftrightarrow (\hat{\theta}, \underline{a})$$

- define a derivative

$$\varphi(\theta) \equiv \ell_{;V}(\theta; y^0) = \frac{\partial}{\partial V(y)} \ell(\theta; y)\Big|_{y=y^0}$$

- a directional derivative on the sample space
- along with $\ell(\theta; y^0)$ the observed log-likelihood function

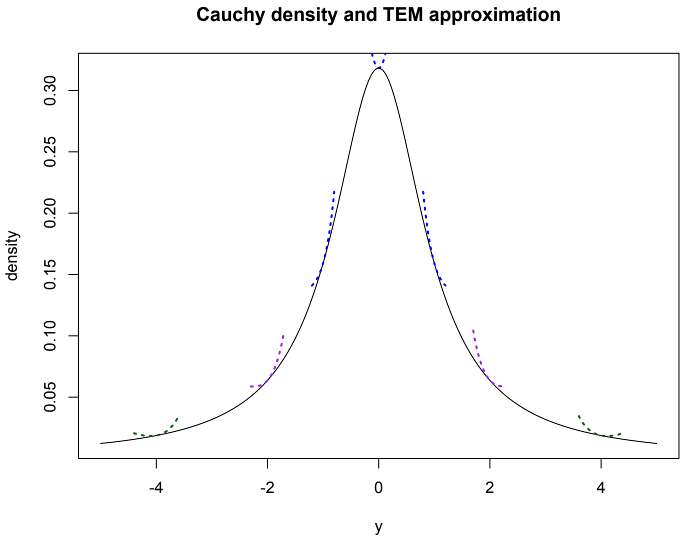
- can be extended to derivative of mean likelihood – usable in wider context Fraser/R Bka 2009

Tangent exponential model

- A continuous model $f(y;\theta)$ on $\mathbb{R}^n$ can be approximated by an exponential family model on $\mathbb{R}^d$:

$$f_{\text{TEM}}(s;\theta)ds = \exp\{\varphi(\theta)'s + \ell^0(\theta)\}h(s)ds \qquad (1)$$

- s is a score variable on $\mathbb{R}^d$ $\qquad s(y) = -\ell_{\varphi}(\hat{\theta}^0; y)$
- $\ell^0(\theta) = \ell(\theta; y^0)$ is the observed log-likelihood function
- $\varphi(\theta) = \varphi(\theta; y^0)$ is the directional derivative $\ell_{;V}(\theta; y^0)$
- (1) approximates original model to $O(n^{-1})$
- gives approximation to the p-value for testing θ
- p-value is accurate to $O(n^{-3/2})$



Cauchy density and TEM approximation
density
0.05
0.10
0.15
0.20
0.25
0.30
-4
-2
0
2
4
y

Example: microscopic fluorescence

- "tracking of microscopic fluorescent particles attached to biological specimens" Hughes et al., AOAS, 2010
- "CCD (charge-coupled device) camera attached to a microscope used to observe the specimens repeatedly"
- "we introduce an improved technique for analyzing such images over time"
- Model for counts:

$$Z_i \sim N(f_i, f_i + \psi), \quad f_i \simeq B + \sum_j A_j \exp\left(-\frac{(x_i - x_j)^2 + (y_i - y_j)^2}{S^2}\right)$$

- f_i developed from a model for photon emission; Normal approximation to Poisson; ψ catches the instrument error

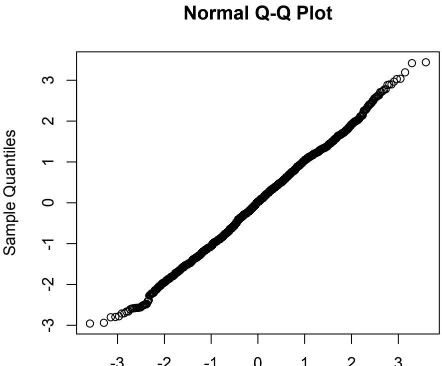
... microscopic fluorescence

- "Our method, which applies maximum likelihood principles, improves the fit to the data, derives accurate standard errors from the data with minimal computation, and uses model-selection criteria to "count" the fluorophores in an image"
- "likelihood ratio tests are used to select the final model"
- potential for improved inference using likelihood methods?

... a simpler model

$$Y_i \sim N(\mu_i, \mu_i + \psi), \quad \mu_i = \exp(\beta_0 + \beta_1 x_i)$$

approximate pivot r^* constructed from $\ell(\theta; y^0), \varphi(\theta; y^0)$ should follow a $N(0,1)$ distribution – simulations



More realistic models

- for example for analytic inferences for survey data
- stochastic processes in space or space-time
- extremes in several dimensions
- frailty models in survival data
- longitudinal data
- family-based genetic data and other forms of clustering
- estimation of recombination rates from SNP data
- ...

Example: Gaussian random field

- scalar output y at $p-$dimensional input $x = (x_1, \ldots, x_p)$
- $$y(x) = \phi(x)^T\beta + Z(x), \quad Z(x) \text{ Gaussian process on } \mathbb{R}^p$$
- $$\text{Cov}\{Z(x_1), Z(x_2)\} = \sigma^2 \prod_{i=1}^{p} R(|x_{1i} - x_{2i}|; \theta)$$
- $$R(|x_{1i} - x_{2i}|) = \exp\{-\gamma_i |x_{1i} - x_{2i}|^\alpha\}$$
- anisotropic covariance matrix for inputs on different scales
- application to computer experiments — Ximing Xu, U Toronto; Derek Bingham, SFU

... Gaussian random field

$\mathbf{y}^n = (y_1, \ldots, y_n) = \{y(x_1), \ldots, y(x_n)\}$, at n locations x_i in $\mathbb{R}^p$

$$\ell(\beta, \sigma, \theta) = -\frac{1}{2}\{n \log \sigma^2 + \log |R(\boldsymbol{\theta})| + \frac{1}{\sigma^2}(\mathbf{y}^n - \boldsymbol{\Phi}\boldsymbol{\beta})^{\mathrm{T}} R^{-1}(\boldsymbol{\theta})(\mathbf{y}^n - \boldsymbol{\Phi}\boldsymbol{\beta})\},$$

computation of R^{-1} is $O(n^3)$, n typically 100s or 1000s

solution – make the correlation matrix sparse

solution – simplify the likelihood function

Example: spatial GLM

- generalized linear geostatistical model

$$E\{Y(x) \mid Z(x)\} = g\{\phi(x)^T\beta + Z(x)\}, x \in \mathbb{R}^2 \text{ or } \mathbb{R}^3$$

- random intercept $Z(x)$ a stationary Gaussian process
- observed at n locations $y(x_i), i = 1, \ldots, n$
- joint density

$$f(y;\theta) = \int_{\mathbb{R}^n} \prod_{i=1}^{n} f(y_i \mid z_i;\theta) f(\mathbf{z};\theta) dz_1 \ldots dz_n$$

- all random effects are correlated
- simulation methods to evaluate integral – MCMC, etc.
- simplify the likelihood function using bivariate integrals

Composite likelihood

- an m-dimensional vector variable Y with model $f(y;\theta)$
- a set of marginal or conditional events $\{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ with associated "sub" log-likelihood

$$\ell_k(\theta; y) = \log f(y \in \mathcal{A}_k) + a(y)$$

- composite log-likelihood

$$\ell_C(\theta; y) = \sum_{k=1}^{K} \ell_k(\theta; y) + a$$

- inference function obtained by pretending sub-models are independent Lindsay, 1988
- a set of non-negative weights $w_1, \ldots, w_k$
- $\ell_C(\theta; y) = \sum_{i=1}^{K} w_k \ell_k(\theta; y)$

... composite likelihood

- Example: pairwise log-likelihood

$$\ell_{pair}(\theta) = \sum_{r=1}^{m} \sum_{s>r} \log f_2(y_r, y_s; \theta)$$

- Example: Besag's pseudo-likelihood

$$\ell_{pseudo}(\theta) = \sum_{r=1}^{m} \log f(y_r \mid \{y_s : y_s \text{ neighbour of } y_r\}; \theta)$$

- Example: Gaussian random field, $\sigma^2 = 1$

$$-\frac{1}{2} \sum_{r=1}^{n-1} \sum_{s=r+1}^{n} \left\{ \log |R_{r,s}| + (\mathbf{y}_{r,s} - \mathbf{\Phi}_{r,s}\beta)^{\mathrm{T}} R_{r,s}^{-1} (\mathbf{y}_{r,s} - \mathbf{\Phi}_{r,s}\beta) \right\},$$

- $\mathbf{y}_{r,s} = (y_r, y_s)$, with 2×2 correlation matrix $R_{r,s}$

Estimation from composite likelihood

- $\ell_C(\theta) = \sum_{k=1}^{K} \ell_k(\theta; y)$
- $U_C(\theta) = \ell'_C(\theta)$ is an unbiased estimating function
- estimate $\hat{\theta}_C$ from $U_C(\hat{\theta}_c) = 0$ is asymptotically normally distributed:

$$\hat{\theta}_C \dot{\sim} N\{\theta, G^{-1}(\theta)\}$$

- asymptotic variance given by Godambe information

$$G(\theta) = \mathrm{E}\{-U'_C(\theta)\}\mathrm{Var}\{U_C(\theta)\}\mathrm{E}\{-U'_C(\theta)\}$$

Inference from composite likelihood

- inference function $\ell_C(\theta)$
- "log-likelihood ratio statistic"

$$w_C(\theta) = 2\{\ell_C(\hat{\theta}_C) - \ell_C(\theta)\}$$

- complicated asymptotic distribution

$$w_C(\theta) \dot{\sim} \sum_{i=1}^{d} \lambda_i \chi^2_{1i}$$

- λ are eigenvalues of $H^{-1}(\theta)G(\theta)$
- $H(\theta) = \mathrm{E}\{-U_C'(\theta)\}$; $G(\theta) = H(\theta)J^{-1}(\theta)H(\theta)$
- rescaling based on score function can restore χ^2_d distribution for w_C Pace, Salvan, Sartori, 2011

Connections to inference from surveys?

- descriptive parameters defined through estimating equation $\sum_{i \in \mathcal{P}} U_i(\theta_{\mathcal{P}}) = 0$
- estimating equation might be motivated by model, e.g. superpopulation model
- "model assisted inference"
- estimating equation from sample $\sum_{i=1}^{n} w_i U_i(\hat{\theta}) = 0$
- for example, $w_i = 1/\pi_i$ or $w_i = 1/(\pi_i q_i)$
- sandwich estimate of variance
- it's all in the weights...

Wei Lin, Changbao Wu

Guidance from composite likelihood?

- in composite likelihood inference, some surprises
- optimal weights may be non-computable
- or even negative — Lindsay, Yi, Sun
- choice of sub-likelihoods needs some care
- in some models including more sub-likelihood terms leads to poorer inference
- in some models including higher dimensional sub-components leads to poorer inference — Ximing Xu
- both choice of weights and choice of component likelihoods needs care

Approximate likelihood inference in survey inference

- example: empirical likelihood for nonparametric models

- $\ell(F) = \sum \log p_i$, with constraints
  $p_i > 0, \sum p_i = 1, \sum p_i y_i = \theta$

- for inference about $\theta = E_F(Y)$, or more generally for parameters defined by estimating functions

- Chen, Sitter, Wu: pseudo-empirical likelihood
- design assisted modelling
- replace $\sum \log p_i$ by $\sum \log p_i w_i$, and constraint by post-stratification such as $\sum_{i=1}^{n} p_i x_i = \bar{X}_{\mathcal{P}}$
- confidence intervals using a profile pseudo-empirical likelihood
- needs adjustment to have asymptotic χ^2 distribution
- rescaling by the design effect

Likelihood for complex models

- Approximate Bayesian Computation
- "an essential tool for the analysis of complex stochastic models" Robert et al. 2011 PNAS
- generate θ' from the prior $\pi(\theta)$
- generate z from the model $p(z \mid \theta')$
- compare $S(z)$ to $S(y)$ using some distance measure $\rho\{S(z), S(y)\}$; if $\rho < \epsilon$ then θ' is a sample from the posterior $\pi(\theta \mid y)$
- actually from $\pi(\theta \mid y, z)$, but this is assume $\approx \pi(\theta \mid y)$
- Robert et al. show that the method can be poor if "$S(\cdot)$ is far from sufficient"
- especially for choosing between models