Likelihood inference in complex settings

Nancy Reid

with Uyen Hoang, Wei Lin, Ximing Xu
Likelihood inference for simple problems

Higher order approximation

Harder problems

Approximations to likelihoods
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
  \[ u(\theta) = \frac{\partial \ell(\theta; y)}{\partial \theta} \]

- 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

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

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

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

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

\[
j(\hat{\theta}) = -\ell''(\hat{\theta})/n \quad \mathcal{I}(\theta) = E\{j(\theta)\}
\]
...approximate pivotals
...approximate pivotal
...approximate pivotals
...approximate pivotals
...approximate pivotals
...approximate pivotals

\[ w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \sim \chi^2_d \]
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)
  \[
  f(\hat{\theta} | a; \theta) = \frac{\exp\{\ell(\theta; y)\}}{\int \exp\{\ell(\theta; y)\} d\theta}
  \]

- \((y_1, \ldots, y_n) \leftrightarrow (\hat{\theta}, a_1, \ldots, a_n) \quad a_i = y_i - \hat{\theta}\)

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

- \(p^*\) approximation:
  \[
p^*(\hat{\theta} | a; \theta) = c(\theta, a)|j(\hat{\theta})|^{1/2} \exp\{\ell(\theta; \hat{\theta}, a) - \ell(\hat{\theta}; \hat{\theta}, a)\}
  \]
A simpler approach

- avoid
  
  \[ (y_1, \ldots, y_n) \leftrightarrow (\hat{\theta}, a) \]

- define a derivative
  
  \[ \varphi(\theta) \equiv \ell; V(\theta; y^0) = \left. \frac{\partial}{\partial V(y)} \ell(\theta; y) \right|_{y=y^0} \]

- 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
Tangent exponential model

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

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

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

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

\[
\ell(\beta, \sigma, \theta) = -\frac{1}{2} \left\{ n \log \sigma^2 + \log |R(\theta)| + \frac{1}{\sigma^2} (y^n - \Phi \beta)^T R^{-1}(\theta) (y^n - \Phi \beta) \right\},
\]

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

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

- 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

\[
\ell_C(\theta; y) = \sum_{i=1}^{K} w_k \ell_k(\theta; y)
\]

Lindsay, 1988
... composite likelihood

- Example: pairwise log-likelihood

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

- Example: Besag’s pseudo-likelihood

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

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

\[-\frac{1}{2} \sum_{r=1}^{n-1} \sum_{s=r+1}^{n} \left\{ \log |R_{r,s}| + (y_{r,s} - \Phi_{r,s}\beta)^T R^{-1}_{r,s} (y_{r,s} - \Phi_{r,s}\beta) \right\}, \]

- \( y_{r,s} = (y_r, y_s) \), with \( 2 \times 2 \) 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:
  \[ \hat{\theta}_C \sim N\{\theta, G^{-1}(\theta)\} \]

- asymptotic variance given by Godambe information
  \[ G(\theta) = \mathbb{E}\{-U_C'(\theta)\} \mathbb{V}\{U_C(\theta)\} \mathbb{E}\{-U_C'(\theta)\} \]
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

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

  \begin{itemize}
  \item $\lambda$ are eigenvalues of $H^{-1}(\theta)G(\theta)$
  \item $H(\theta) = \mathbb{E}\{-U'_C(\theta)\}$; $G(\theta) = H(\theta)J^{-1}(\theta)H(\theta)$
  \item rescaling based on score function can restore $\chi^2_d$ distribution for $w_C$
  \end{itemize}

Pace, Salvan, Sartori, 2011
Connections to inference from surveys?

- descriptive parameters defined through estimating equation \( \sum_{i \in P} U_i(\theta_P) = 0 \)

- 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
- 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
- both choice of weights and choice of component likelihoods needs care

Lindsay, Yi, Sun

Ximing Xu
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 \( \sum_{i=1}^{n} p_i x_i = \bar{X}_P \)
- confidence intervals using a profile pseudo-empirical likelihood
- needs adjustment to have asymptotic \( \chi^2 \) distribution
- rescaling by the design effect
Likelihood inference for simple problems

Higher order approximation

Harder problems

Approximations to likelihoods

Likelihood for complex models

• Approximate Bayesian Computation
• “an essential tool for the analysis of complex stochastic models”
  Robert et al. 2011 PNAS
• generate $\theta'$ 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 $\theta'$ 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