Aspects of Likelihood Inference

Nancy Reid

October 16, 2013

JACOBI BERNOULLI,

ARS CONJECTANDI,

OPUS POSTHUMUM.

TRACTATUS

DE SERIEBUS INFINITIS,

DE LUDO PILÆ

RETICULARIS.

BASILEÆ,

Impensis THURNISIORUM, Fratrum.

Models and likelihood

- Model for the probability distribution of y given x
- Density $f(y \mid x)$ with respect to, e.g., Lebesgue measure
- Parameters for the density $f(y \mid x; \theta)$, $\qquad \theta = (\theta_1, \ldots, \theta_d)$

- Likelihood function $L(\theta; y^0) \propto f(y^0; \theta)$

- often $\theta = (\psi, \lambda)$
- θ could have very large dimension, $d > n$

  typically $y = (y_1, \ldots, y_n)$
- θ could have infinite dimension $E(y \mid x) = \theta(x)$ 'smooth', in principle

Why likelihood?

- makes probability modelling central
- emphasizes the inverse problem of reasoning from y^0 to θ or $f(\cdot)$
- suggested by Fisher as a measure of plausibility

Royall, 1994

$$
\begin{array}{ll}
L(\hat{\theta})/L(\theta) \in (1,3) & \text{very plausible;} \\
L(\hat{\theta})/L(\theta) \in (3,10) & \text{implausible;} \\
L(\hat{\theta})/L(\theta) \in (10,\infty) & \text{very implausible}
\end{array}
$$

- converts a 'prior' probability $\pi(\theta)$ to a posterior $\pi(\theta \mid y)$ via Bayes' formula

- provides a conventional set of summary quantities for inference based on properties of the postulated model

Widely used

ELSEVIER

Cold Regions Science and Technology

Available online 4 October 2013

In Press, Accepted Manuscript — Note to users

A Generalized Probabilistic Model of Ice Load Peaks on Ship Hulls in Broken-Ice Fields

A. Suyuthi[a], B.J. Leira[a], K. Riska[b, c]

[a] Department of Marine Technology, NTNU, Trondheim, Norway

[b] Centre of Ships and Offshore Structures (CeSOS), Trondheim, Norway

[c] ILS OY, Helsinki, Finland

... widely used

ELSEVIER

Molecular Phylogenetics and Evolution

Available online 3 October 2013

In Press, Uncorrected Proof — Note to users

Diversification of *Scrophularia* (Scrophulariaceae) in the Western Mediterranean and Macaronesia – Phylogenetic relationships, reticulate evolution and biogeographic patterns

Agnes Scheunert, Günther Heubl

Systematic Botany and Mycology, Department Biology I, Ludwig-Maximilians-University, GeoBio Center LMU, Menzinger Strasse 67, 80638 Munich, Germany

... widely used

Empirical growth curve estimation considering multiple seasonal compensatory growths of body weights in Japanese Thoroughbred colt
and fillies.
(PMID:24085406)

Abstract | Citations | BioEntities | Related Articles | External Links

Onoda T, Yamamoto R, Sawmura K, Inoue Y, Murase H, Nambo Y, Tozaki T, Matsui A, Miyake T, Hirai N
Comparative Agricultural Sciences, Graduate School of Agriculture, Kyoto University, Kyoto 606-8502, Japan
Journal of Animal Science [2013]

Type: Journal Article

... widely used

ACADEMY PUBLISHER

Journal of Networks

HOME LOG IN REGISTER SEARCH CURRENT ARCHIVES

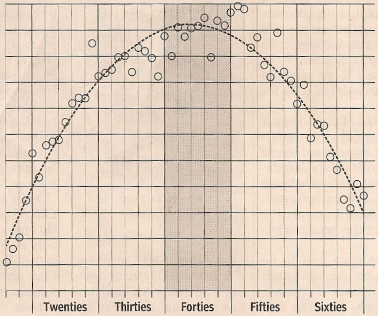
Home > Vol 8, No 10 (2013) > **Wang**

Journal of Networks, Vol 8, No 10 (2013), 2220-2226, Oct 2013
doi:10.4304/jnw.8.10.2220-2226

Low-Complexity Carrier Frequency Offset Estimation Algorithm in TD-LTE

Dan Wang, Weiping Shi, Xiaowen Li

... widely used


HAVING A MID-LIFE CRISIS? YOU'RE NOT ALONE

A study involving two million people in 72 countries found men and women were less happy in their 40s but that improved in later life.

SOURCES: IS WELL-BEING U-SHAPED OVER THE LIFE CYCLE?

RICHARD JOHNSON / NATIONAL POST

National Post, Toronto, Jan 30 2008

... why likelihood?

- likelihood function depends on data only through sufficient statistics
- "likelihood map is sufficient" Fraser & Naderi, 2006

- gives exact inference in transformation models
- "likelihood function as pivotal" Hinkley, 1980

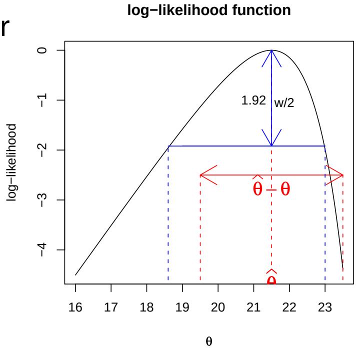
- provides summary statistics with known limiting distribution
- leading to approximate pivotal functions, based on normal distribution
- likelihood function $+$ sample space derivative gives better approximate inference

Derived quantities

- maximum likelihood estimator
  $\hat{\theta} = \arg\sup_\theta \log L(\theta; y)$
  $= \arg\sup_\theta \ell(\theta; y)$

- observed Fisher information
  $j(\hat{\theta}) = -\partial^2 \ell(\theta)/\partial\theta^2$

- efficient score function
  $\ell'(\theta) = \partial\ell(\theta; y)/\partial\theta$



$\ell'(\hat{\theta}) = 0$ assuming enough regularity

- $\ell'(\theta; y) = \sum_{i=1}^n \log f_{Y_i}(y_i; \theta)$, $\qquad y_1, \ldots, y_n$ independent

Approximate pivots

scalar parameter of interest

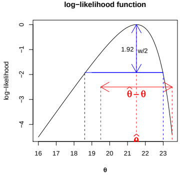
- profile log-likelihood $\ell_p(\psi) = \ell(\psi, \hat\lambda_\psi)$
- $\theta = (\psi, \lambda)$; $\hat\lambda_\psi$ constrained maximum likelihood estimator

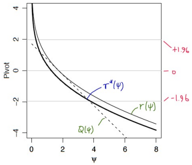
$$r_e(\psi; y) = (\hat\psi - \psi) j_p^{1/2}(\hat\psi) \quad \dot\sim \quad N(0,1)$$

$$r(\psi; y) = \pm\sqrt{[2\{\ell_p(\hat\psi) - \ell_p(\psi)\}]} \quad \dot\sim \quad N(0,1)$$

$$\pi_m(\psi \mid y) \quad \dot\sim \quad N\{\hat\psi, j_p^{-1/2}(\hat\psi)\}$$

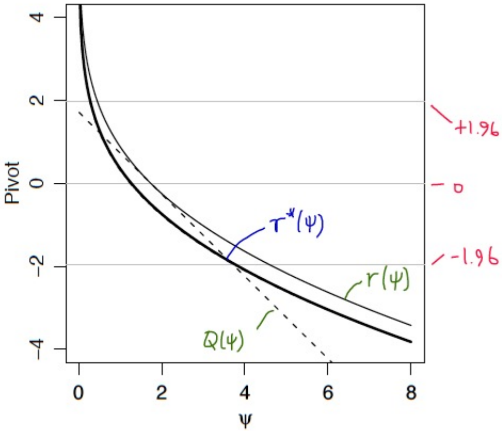
$j_p(\psi) = -\ell_p''(\psi)$; profile information





... approximate pivots

scalar parameter of interest



... approximate pivots

scalar parameter of interest

- profile log-likelihood $\ell_p(\psi) = \ell(\psi, \hat\lambda_\psi)$
- $\theta = (\psi, \lambda)$; $\hat\lambda_\psi$ constrained maximum likelihood estimator

$$r_e(\psi; y) = (\hat\psi - \psi) j_p^{1/2}(\hat\psi) \quad \dot\sim \quad N(0,1)$$

$$r(\psi; y) = \pm\sqrt{[2\{\ell_p(\hat\psi) - \ell_p(\psi)\}]} \quad \dot\sim \quad N(0,1)$$

$$\pi_m(\psi \mid y) \quad \dot\sim \quad N\{\hat\psi, j_p^{-1/2}(\hat\psi)\}$$

$$r^*(\psi; y) = r(\psi) + \frac{1}{r(\psi)} \log\left\{\frac{Q_F(\psi)}{r(\psi)}\right\} \quad \dot\sim \quad N(0,1)$$

$$r_B^*(\psi; y) = r(\psi) + \frac{1}{r(\psi)} \log\left\{\frac{Q_B(\psi)}{r(\psi)}\right\} \quad \dot\sim \quad N(0,1)$$

The problem with profiling

- $\ell_{\mathrm{p}}(\psi) = \ell(\psi, \hat{\lambda}_\psi)$ used as a 'regular' likelihood, with the usual asymptotics
- neglects errors in the estimation of the nuisance parameter
- can be very large when there are many nuisance parameters

- example: normal theory linear regression $\hat{\sigma}^2 = RSS/n$ usual estimator $RSS/(n-k)$ k the number of regression coefficients
- badly biased if k large relative to n
- inconsistent for σ^2 if $k \to \infty$ with n fixed
- example fitting of smooth functions with large numbers of spline coefficients

Conditional and marginal likelihoods

$$f(y;\psi,\lambda) \propto f_1(s \mid t;\psi)f_2(t;\lambda)$$

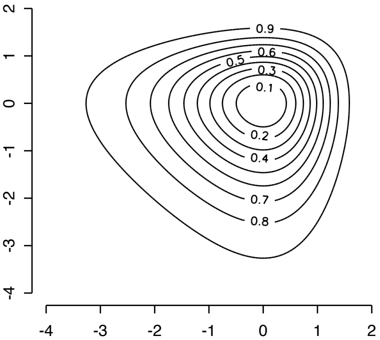
- $L(\psi,\lambda) \propto L_c(\psi)L_m(\lambda)$, where L_1 and L_2 are genuine likelihoods, i.e. proportional to genuine density functions
- $L_p(\psi)$ is a conditional likelihood $L_c(\psi)$, and estimation of λ has no impact on asymptotic properties
- s is conditionally sufficient , t is marginally ancillary, for ψ

- hardly ever get so lucky
- but might expect something like this to hold approximately, which it does, and this is implemented in r_F^* formula
  automatically Brazzale, Davison, R 2007

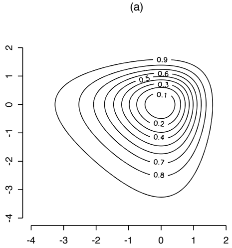
Directional inference

- vector parameter of interest $\theta = (\underline{\psi}, \underline{\lambda}), \psi \in \mathbb{R}^q$
- approximate pivotal quantity
  $w(\psi) = 2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\} \dot{\sim} \chi^2_q$

(a)



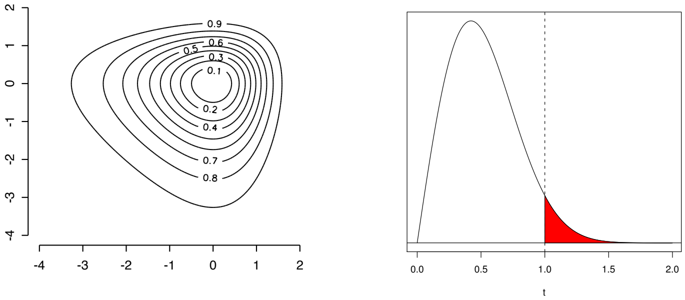
... directional inference



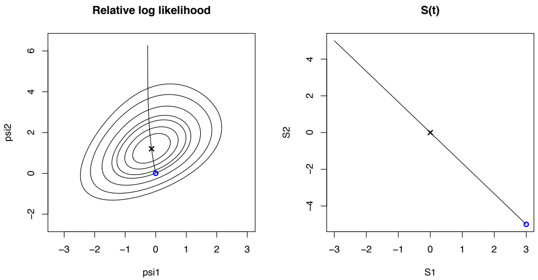
$$2\{\ell_{\mathrm{p}}(\hat{\psi}) - \ell_{\mathrm{p}}(\psi)\}/\{1 + B(\psi)/n\} \mathrel{\dot{\sim}} \chi^2_q$$

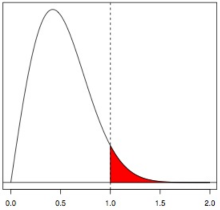
... directional inference

(a)



... directional tests $\mathcal{L}^* = ts^0 + (1-t)s_\psi$





○ null hypothesis of independence $t = 0$

X observed value of s $t = 1$

$$p\text{-value} = \frac{\int_1^\infty t^{d-1} g\{s(t);\psi\}dt}{\int_0^\infty t^{d-1} g\{s(t);\psi\}dt}$$

like a 2-sided p-value

Pr (response $>$ observed $\mid$ response > 0) Davison et al. 2014

Model selection/choice

- likelihood inference very/completely dependent on correctness of assumed model
- role in model choice?

- nested models:
  - log-likelihood ratio $w = 2\{\ell_p(\hat{\psi}) - \ell_p(\psi = 0)\}$
  - assess consistency of data with $\psi = 0$, i.e. with simpler model
    with either usual asymptotics or higher order versions

- if models are non-nested, for example log-normal vs gamma, then a different asymptotic theory is needed

  separate families, Cox 1961,2, 2013

... model selection/choice

- from prediction in time series,

$$AIC = -2\log L(\hat{\theta}; y) + 2d$$

- from model choice in Bayesian inference, combined with Laplace approximation

$$BIC = -2\log L(\hat{\theta}; y) + \log(n)d$$

- relative values of interest only, in models of differing dimensions

- a ‘non-likelihood’ approach $f(y;\theta) \propto f_m(s;\theta)f_c(t \mid s)$; second component can be used for a test of model fit

Extending the likelihood function

- asymptotic results provide some theoretical insight
- often difficult to apply in complex models, especially models with complex dependencies
- is likelihood inference still relevant in more complex settings?

- inference based on the likelihood function provides a standard set of tools
- “we believe that greater use of the likelihood based approaches and goodness-of-fit measures can help improve the quality of neuroscience data analysis”

  Brown et al.

- one way to make models more complex is to add more parameters
- although we’ve seen that this can lead to difficulties

... extending likelihood inference

- various inference functions have been proposed
- typically in the context of particular applications or model classes
- with a bewildering number of names: quasi-likelihood, h-likelihood, penalized quasi-likelihood, pseudo-likelihood, composite likelihood, partial likelihood, empirical likelihood
- to name a few
- why so many choices?
- hope to get summary statistics with reasonable properties
- hope that the inference function itself will carry some information
- in some cases hope to combine these functions with a prior probability to simplify Bayesian computations

Pocket guide to other likelihoods

- introduce dependence through latent random variables
- probability model then involves integrating over their distribution
- only analytically possible is special cases
- Laplace approximation to this integral is called penalized quasi-likelihood Breslow & Clayton, 1993
- If $g\{E(y)\} = X\theta + Zb$, then leads to

$$\ell(\theta, b; y) - \frac{1}{2} b^T D^{-1}(\theta) b$$

- the derivation generalizes the quasi-likellihood used in GLMs, which specify mean and variance functions only
- combining marginal likelihoods for dispersion parameters with GLMMs leads to *h*-likelihood

Nelder & Lee 1996

... pocket guide

- Composite likelihood, also called pseudo-likelihood

  Besag, 1975
- reduce high-dimensional dependencies by ignoring them

- for example, replace $f(y_1, \ldots, y_k; \theta)$ by

$$\text{pairwise marginal} \qquad \prod_{j<j'} f_2(y_j, y_{j'}; \theta), \quad \text{or}$$

$$\text{conditional} \qquad \prod_{j} f_c(y_j \mid y_{\mathcal{N}(j)}; \theta)$$

- a type of modelling robustness
- limit theorems related to mis-specified models

$$\hat{\theta}_{CL} \dot{\sim} N\{\theta, G^{-1}(\theta)\}, \quad G(\theta) = H(\theta)J^{-1}(\theta)H(\theta)$$

$$J(\theta) = \text{var}\ell'_{CL}(\theta), \quad J(\theta) = -\text{E}\ell''_{CL}(\theta)$$

... pocket guide

- Semi-parametric models, leading to partial likelihood
- e.g. proportional hazards model for survival data

  Cox 1972, 1975
- partial likelihood has the usual asymptotic properties of profile likelihood Murphy and Van der Waart, 2000
- obtained via a projection argument of the score function for the parameter of interest
- e.g. partially linear regression models, with 'smooth' function replaced by a linear combination of basis functions

$$E(y_i) = \beta_0 + \beta_1 x_i + \sum_{j=1}^{J} \gamma_j B(z_i)$$

- maximize a penalized log-likelihood function $\ell(\beta, \gamma) + \lambda p(\gamma)$

  Fan & Li, 2001; Green, 1987; Van der Vaart (1998, Ch. 25)

Simulated likelihoods/posteriors

- Approximate Bayesian Computation
  - simulate θ' from $\pi(\theta)$
  - simulate data z from $f(\cdot;\theta)$
  - if $z = y$ then θ' is an observation from $\pi(\theta \mid y)$
  - actually $s(z) = s(y)$ for some set of statistics
  - actually $\rho\{s(z), s(y)\} < \epsilon$ for some distance function $\rho(\cdot)$
- related to simulation by MCMC for computation of MLEs

  Geyer & Thompson
- can be used for approximate construction of likelihood
- and is related to generalized method of moments

  Cox & Kartsonakis, 2012

Conclusion

- likelihood inference is really model-based inference
- models are important for most scientific work
- important to understand their implications and limitations
- and to use them as efficiently as possible
- with or without 'Big Data'


JACOBI BERNOULLI,

ARS CONJECTANDI,

OPUS POSTHUMUM.

TRACTATUS

DE SERIEBUS INFINITIS,

DE LUDO PILÆ

RETICULARIS.

BASILEÆ,

Impensis THURNISIORUM, Fratrum.