Approximate Likelihoods

Nancy Reid
University of Toronto

August 14, 2015

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 $f(y \mid x; \theta)$, $\theta = (\theta_1, \dots, \theta_d)$
- Data $y = (y_1, \dots, y_n)$ often independent

- Likelihood function $L(\theta; y) \propto f(y; \theta)$ $(y_1, \dots, y_n)$
- log-likelihood function $\ell(\theta; y) = \log L(\theta; y)$

- often $\theta = (\psi, \lambda)$

- θ could have very large dimension, $d > n$

- θ could have infinite dimension in principle

  $E(y \mid x) = \theta(x)$ 'smooth'

Why likelihood?

- makes probability modelling central $\ell(\theta; y) = \log f(y; \theta)$
- emphasizes the inverse problem of reasoning $y \to \theta$
- converts a 'prior' probability to a posterior $\pi(\theta) \to \pi(\theta \mid y)$
- provides a conventional set of summary quantities: maximum likelihood estimator, score function, ...
- provides summary statistics with known limiting distribution
- these define approximate pivotal quantities, based on normal distribution
- basis for comparison of models, using AIC or BIC

Widely used

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

▹ PP-A09-12
A Semiparametric Empirical Likelihood on the Linear Models with Covariates Parametrically Transformed
Zhang, Jing Hua — Beijing Univ. of Tech.
Xue, Liugen — Beijing Univ. of Tech.

Screen Shot 2015-08-11 at 11.32.44 PM

▹ PP-A09-17
Empirical Likelihood in Generalized Linear Models for Longitudinal Data with Dropout

Screen Shot 2015-08-11 at 11.32.54 PM

Guo, Donglin — Beijing Univ. of Tech.
Xue, Liugen — Beijing Univ. of Tech.

Screen Shot 2015-08-11 at 11.32.58 PM

▹ PP-A09-8
Generalized Empirical Likelihood Inference for Longitudinal Data with Missing Response Variables and Error-Prone Covariates
Liu, Juanfang — Beijing Univ. of Tech.
Xue, Liugen — Beijing Univ. of Tech.

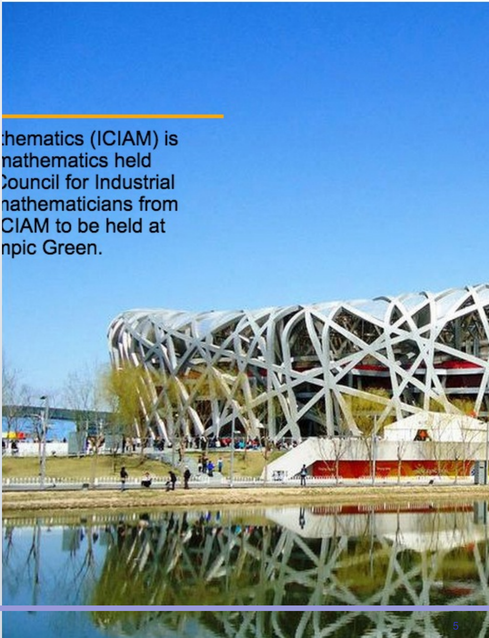
Screen Shot 2015-08-11 at 11.32.34 PM

▶ MS-Fr-D-48-3 — 14:30–15:00
Image Reconstruction and Interpretation in Positron Emission Tomography for Small Animals (micro-PET)
Garbarino, Sara — Department of Mathematics, Univ. of Genoa

Screen Shot 2015-08-11 at 11.32.19 PM

▶ MS-Fr-D-36-2 — 14:00–14:30
A Randomized Likelihood Method for Data Reduction in Large-scale Inverse Problems

Screen Shot 2015-08-11 at 11.32.06 PM

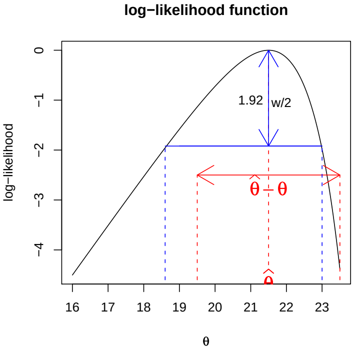


... why likelihood?

- provides a conventional set of summary quantities: maximum likelihood estimator, score function, ...
- provides summary statistics with known limiting distribution

Important summaries

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



- observed Fisher information

$$j(\hat{\theta}) = -\left.\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right|_{\hat{\theta}}$$

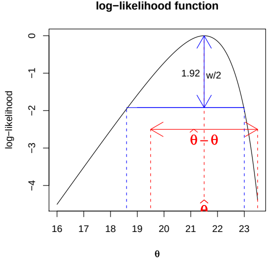
- efficient score function
  $\ell'(\theta) = \partial \ell(\theta; y)/\partial\theta$

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

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

Limit theorems

- $\ell'(\theta) j^{-1/2}(\hat\theta) \xrightarrow{\mathcal{L}} N(0,1)$
- $(\hat\theta - \theta) j^{1/2}(\hat\theta) \xrightarrow{\mathcal{L}} N(0,1)$
- $2\{\ell(\hat\theta) - \ell(\theta)\} \xrightarrow{\mathcal{L}} \chi^2_1$
- under the model $f(y;\theta)$
  regularity conditions



- approximate pivots

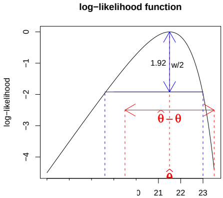
$$r_e(\theta) = (\hat\theta - \theta) j^{1/2}(\hat\theta)$$

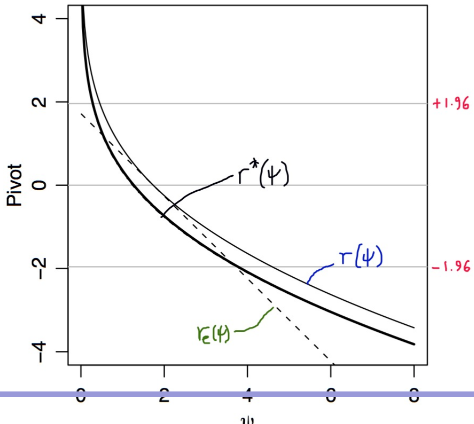
$$r(\theta) = \pm\sqrt{[2\{\ell(\hat\theta) - \ell(\theta)\}]}$$

... approximate pivots

$$r_e(\theta) = (\hat{\theta} - \theta) j^{1/2}(\hat{\theta})$$

$$r(\theta) = \pm\sqrt{[2\{\ell(\hat{\theta}) - \ell(\theta)\}]}$$





Complicated likelihoods

generalized linear mixed models

GLM: $y_{ij} \mid u_i \sim \exp\{y_{ij}\eta_{ij} - b(\eta_{ij}) + c(y_{ij})\}$

linear predictor: $\eta_{ij} = x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}}u_i \qquad j{=}1,\ldots n_i; \quad i{=}1,\ldots m$

random effects: $u_i \sim N_k(0, \Sigma)$

log-likelihood:

$$\ell(\beta,\Sigma) = \sum_{i=1}^{m}\Bigg(y_i^{\mathrm{T}}X_i\beta - \frac{1}{2}\log|\Sigma| + \log\int_{\mathbb{R}^k}\exp\{y_i^{\mathrm{T}}Z_iu_i - 1_i^{\mathrm{T}}b(X_i\beta + Z_iu_i) - \frac{1}{2}u_i^{\mathrm{T}}\Sigma^{-1}u_i\}du_i\Bigg)$$

Ormerod & Wand 2012

... complicated likelihoods

multivariate extremes: example, wind speed at d locations

vector observations: $(X_{1i}, \ldots, X_{di}),\ i = 1, \ldots, n$

component-wise maxima: $Z_1, \ldots, Z_d; Z_j = \max(X_{j1}, \ldots, X_{jn})$

Z_j are transformed (centered and scaled)

joint distribution function:

$$\Pr(Z_1 \le z_1, \ldots, Z_d \le z_d) = \exp\{-V(z_1, \ldots, z_d)\}$$

$V(\cdot)$ can be parameterized via Gaussian process models

likelihood : need the joint derivatives of $V(\cdot)$

combinatorial explosion

Davison et al., 2012

... complicated likelihoods

Ising model:

$$f(y;\theta) = \exp\Big(\sum_{(j,k)\in E} \theta_{jk} y_j y_k\Big)\frac{1}{Z(\theta)} \qquad j,k = 1,\dots,K$$

observations: $y_i = \pm 1$; binary property of a node i in a graph with K nodes

parameter: θ_{jk} measures strength of interaction between nodes i and j

E is the set of edges between nodes

partition function:

$$Z(\theta) = \sum_y \exp\Big(\sum_{(j,k)\in E} \theta_{jk} y_j y_k\Big)$$

Davison 2000 §6.2; Ravikumar et al. (2010); Xue et al. (2012)

... complicated likelihoods

$M/G/1$ queue: exponential arrival times, general service times, single server

observations y_i: times between departures from the queue

unobserved variables V_i: arrival time of customer i

model:

- $V_1 \sim \text{Exp}(\theta_3)$
- $V_i \mid V_{i-1} \sim V_{i-1} + \text{Exp}(\theta_3)$
- $Y_i \mid X_{i-1}, V_i \sim \text{Uniform}\{\theta_1 + max(0, V_i - X_{i-1}),$
  $\theta_2 + max(0, V_i - X_{i-1})\}$ $\quad X_i = \sum_{j=1}^{i} Y_j \quad G = U(\theta_1, \theta_2)$

Likelihood

$$L(\theta; y) = \int \cdots \int f(v_1 \mid \theta) \prod_{i=1}^{n} f(v_i \mid v_{i-1}, \theta) \prod_{i=1}^{n} f(y_i \mid v_i, x_{i-1}, \theta) dv_1 \cdots dv_n$$

Heggland & Frigessi, 2004

Fearnhead & Prangle, 2012

What's a poor statistician to do?

- simplify the likelihood
  - composite likelihood
  - variational approximation
  - Laplace approximation to integrals

- change the mode of inference
  - quasi-likelihood
  - indirect inference

- simulate
  - approximate Bayesian computation
  - Markov chain Monte Carlo

Composite likelihood

- also called pseudo-likelihood Besag, 1975
- reduce high-dimensional dependencies by ignoring them

- for example, replace $f(y_{i1}, \ldots, y_{ik}; \theta)$ by

pairwise marginal $\prod_{j<j'} f_2(y_{ij}, y_{ij'}; \theta)$, or

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

- Composite likelihood function

$$CL(\theta; y) \propto \prod_{i=1}^{n} \prod_{j<j'} f_2(y_{ij}, y_{ij'}; \theta)$$

- Composite ML estimates are consistent, asymptotically normal, not fully efficient Lindsay, 1988; Varin R Firth, 2011

Example: spatial extremes

Davison et al., 2012; & Huser, 2015

$$\Pr(Z_1 \leq z_1, \ldots, Z_d \leq z_d) = \exp\{-V(z_1, \ldots, z_d; \theta)\}$$

- pairwise composite likelihood used to compare the fits of several competing models
- model choice using "CLIC", an analogue of AIC

  $-2\log(\widehat{CL}) + \text{tr}(J^{-1}K)$

- Davison et al. 2012 applied this to annual maximum rainfall at several stations near Zurich
- "fitting max-stable processes to spatial or spatio-temporal block maxima is awkward ... the use of composite likelihoods ... has become widely used" Davison & Huser

Example: Ising model

Ising model:

$$f(y;\theta) = \exp\Big(\sum_{(j,k)\in E} \theta_{jk} y_j y_k\Big) \frac{1}{Z(\theta)} \qquad j,k = 1,\ldots,K$$

neighbourhood contributions

$$f(y_j \mid y_{(-j)};\theta) = \frac{\exp(2y_j \sum_{k\neq j} \theta_{jk} y_k)}{\exp(2y_j \sum_{k\neq j} \theta_{jk} y_k) + 1} \quad = \quad \exp \ell_j(\theta; y)$$

penalized CL estimation based on sample $y^{(1)},\ldots,y^{(n)}$

$$\max_\theta \left\{ \sum_{i=1}^{n} \sum_{j=1}^{K} \ell_j(\theta; y^{(i)}) - \sum_{j<k} P_\lambda(|\theta_{jk}|) \right\}$$

Xue et al., 2012

Ravikumar et al., 2010

Variational methods

Ormerod & Wand, 2010

- in a Bayesian context, want $f(\beta \mid y)$
  use an approximation $q(\beta)$
- dependence of q on y suppressed

- choose $q(\beta)$ to be
  - simple to calculate
  - close to posterior

- simple to calculate
  - $q(\beta) = \prod q_j(\beta_j)$
  - simple parametric family

- close to posterior: miminize Kullback-Leibler divergence between $q(\cdot)$ and $f(\cdot \mid y)$

... variational methods

Titterington, 2006

- close to posterior: miminize Kullback-Leibler divergence

$$KL(q \,||\, f_{post}) = \int q(\beta)\log\{q(\beta)/f(\beta \mid y)\}d\beta$$

- equivalent to

$$\max_q \int q(\beta)\log\{f(y,\beta)/q(\beta)\}d\beta$$

- because

$$\log f(y;\theta) \geq \int q(\beta)\log\{f(y,\beta;\theta)/q(\beta)\}d\beta$$

- in a likelihood context

$$\log f(y;\theta) = \log \int f(y \mid \beta;\theta)f(\beta)d\beta$$

here β represent random effects u, or b, or ...

Example: GLMM

Ormerod & Wand, 2012

log-likelihood:

$$\ell(\beta, \Sigma) = \sum_{i=1}^{m} \Big(y_i^{\mathrm{T}} X_i \beta - \frac{1}{2} \log |\Sigma| + \log \int_{\mathbb{R}^k} \exp\{ y_i^{\mathrm{T}} Z_i u_i - 1_i^{\mathrm{T}} b(X_i\beta + Z_i u_i) - \frac{1}{2} u_i^{\mathrm{T}} \Sigma^{-1} u_i \} du_i \Big)$$

variational approx:

$$\ell(\beta, \Sigma) \geq \sum_{i=1}^{m} \Big(y_i^{\mathrm{T}} X_i \beta - \frac{1}{2} \log |\Sigma| \Big) + \sum_{i=1}^{m} E_{u \sim N(\mu_i, \Lambda_i)} \Big(y_i^{\mathrm{T}} Z_i u - 1_i^{\mathrm{T}} b(X_i\beta + Z_i u) - \frac{1}{2} u^{\mathrm{T}} \Sigma^{-1} u - \log\{ \phi_{\Lambda_i}(u - \mu_i) \} \Big)$$

simplifies to k one-dim. integrals

... variational approximations

Ormerod & Wand, 2012

- $$\ell(\beta,\Sigma) \geq \ell(\beta,\Sigma,\mu,\Lambda)$$
- variational estimate:

$$\ell(\tilde\beta,\tilde\Sigma,\tilde\mu,\tilde\Lambda) = \arg\max_{\beta,\Sigma,\mu,\Lambda}\ell(\tilde\beta,\tilde\Sigma,\tilde\mu,\tilde\Lambda)$$

- inference for $\tilde\beta,\tilde\Sigma$? consistency? asymptotic normality?

  Hall, Ormerod, Wand, 2011; Hall et al. 2011
- emphasis on algorithms and model selection

  e.g. Tan & Nott, 2013, 2014

- VL: approx $L(\theta;y)$ by a simpler function of θ, e.g. $\prod q_j(\theta)$
- CL: approx $f(y;\theta)$ by a simpler function of y, e.g. $\prod f(y_j;\theta)$

Some Links between Variational Approximation and Composite Likelihoods?

S. Robin

UMR 518 AgroParisTech / INRA Applied Math & Comput. Sc.

MSTGA, Paris, November 22-23, 2012

http://carlit.toulouse.inra.fr/AIGM/pub/Reunion_nov2012/MSTGA-1211-Robin.pdf

Zhang & Schneider 2012 JMLR V22; Grosse 2015 ICML

Indirect inference

- composite likelihood estimator solves $(\partial/\partial\theta)\log CL(\theta; y) = 0$
- solution converges to the true value under conditions ...
- because $\mathrm{E}\{(\partial/\partial\theta)\log CL(\theta; y)\} = 0$

- what happens if an estimating equation $g(y;\theta)$ is biased?
- $g(y_1, \dots, y_n; \tilde{\theta}_n) = 0; \qquad \tilde{\theta}_n \to \theta^*$ $\qquad \mathrm{E}g(Y;\theta^*) = 0$

- $\theta^* = \tilde{k}(\theta)$; invertible? $\theta = k(\theta^*)$ $\qquad \tilde{k}^{-1} \equiv k$

- new estimator $\hat{\theta}_n = k(\tilde{\theta}_n)$
- $k(\cdot)$ is a bridge function, connecting wrong value of θ to the right one Yi & R, 2010; Jiang & Turnbull, 2004

... indirect inference

Smith, 2008; Shalizi, 2013

- model of interest

$$y_t = G_t(y_{t-1}, x_t, \epsilon_t; \theta), \quad \theta \in \mathbb{R}^d$$

- likelihood is not computable, but
  we can simulate from the model

- simple (wrong) model

$$y_t \sim f(y_t \mid y_{t-1}, x_t; \theta^*), \quad \theta^* \in \mathbb{R}^p$$

- find the MLE in the simple model, $\hat{\theta}^* = \hat{\theta}^*(y_1, \ldots, y_n)$, say

- simulate from model of interest for some value θ, compute a new MLE in simple model
- 'good' values of θ give data that reproduces $\hat{\theta}^*$

... indirect inference

Smith, 2008

- simulate samples $y_t^m, \quad m = 1, \ldots, M$ at some value θ
- compute $\hat{\theta}^*(\theta)$ from the simulated data

$$\hat{\theta}^*(\theta) = \arg\max_{\theta^*} \sum_m \sum_t \log f(y_t^m \mid y_{t-1}^m, x_t; \theta^*)$$

- choose θ so that $\hat{\theta}^*(\theta)$ is as close as possible to $\hat{\theta}^*$
- if both model parameters have the same dimension simply invert the 'bridge function'
- usually not, so minimize some measure of distance between $\hat{\theta}(\beta)$ and $\hat{\theta}$
- estimates of θ are consistent, asymptotically normal, but not efficient

Approximate Bayesian Computation

Marin et al., 2010

- simulate θ from prior density $\pi(\cdot)$
- simulate data y' from $f(\cdot; \theta)$
- if $y' = y$ then θ is an observation from posterior $\pi(\cdot \mid y)$
- actually $s(y') = s(y)$ for some set of statistics
- actually $\rho\{s(y'), s(y)\} < \epsilon$ for some distance function $\rho(\cdot)$

Fearnhead & Prangle, 2011

- many variations, using different MCMC methods to select candidate values θ

... approximate Bayesian computation

$M/G/1$ queue: exponential arrival times, general service times, single server

observations y_i: times between departures from the queue

unobserved variables V_i: arrival time of customer i

model:

- $V_1 \sim \text{Exp}(\theta_3)$
- $V_i \mid V_{i-1} \sim V_{i-1} + \text{Exp}(\theta_3)$
- $Y_i \mid X_{i-1}, V_i \sim \text{Uniform}\{\theta_1 + max(0, V_i - X_{i-1}),$
  $\theta_2 + max(0, V_i - X_{i-1})\} \qquad X_i = \sum_{j=1}^{i} Y_j$
- service time $\sim U(\theta_1, \theta_2)$

ABC: use quantiles of departure times as summary statistics

Indirect Inference: use $\bar{y}$, $y_{(1)}$, $\hat{\theta}_2$ from steady-state model

Heggland & Frigessi, 2004

Table 7. Mean quadratic losses for various analyses of 50 $M/G/1$ data sets†

Method	θ_1	θ_2	θ_3
Comparison	1.1	2.2	0.0013
Comparison + regression	*0.020*	1.1	*0.0013*
Semi-automatic ABC	*0.022*	1.0	*0.0013*
Semi-automatic predictors	0.024	1.2	0.0017
Indirect inference	0.18	*0.42*	0.0033

†Losses within 10% of the smallest values for that parameter are italicized.

Fearnhead & Prangle, 2011

ABC and Indirect Inference

Cox & Kartsonaki, 2012

- both methods need a set of parameter values from which to simulate: θ' or θ
- both methods need a set of auxiliary functions of the data $s(y)$ or $\hat{\theta}^*(y)$

- in indirect inference, $\hat{\theta}^*$ is the 'bridge' to the parameters of real interest, θ

- C & K use orthogonal designs based on Hadamard matrices to chose θ'
- and calculate summary statistics focussed on individual components of θ

What's a poor statistician to do?

- simplify the likelihood
  - composite likelihood
  - variational approximation
  - Laplace approximation to integrals
- change the mode of inference
  - quasi-likelihood
  - indirect inference
- simulate
  - approximate Bayesian computation
  - MCMC

Summary

so much to do, so little time!

Summary

- empirical likelihood, weighted likelihood, local likelihood, sieve likelihood, simulated likelihood, ...
- likelihood provides a common set of tools:
  - summary statistics
  - e.g. point estimates and estimates of precision
  - comparison of models
- likelihood puts modelling first
- likelihood puts inference first
- contrast with ‘black-box’ predictions

Thank You!