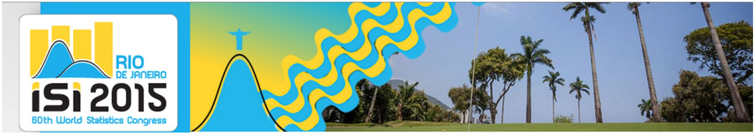
Approximate Likelihoods

Nancy Reid

July 28, 2015



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, ...
- these define approximate pivotal quantities, based on normal distribution

- basis for comparison of models, using AIC or BIC

Example 1: GLMM

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:

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

Ormerod & Wand 2012

Example 2: Poisson AR

Poisson $$f(y_t \mid \alpha_t; \theta) = \exp(y_t \log \mu_t - \mu_t)/y_t!$$

$$\log \mu_t = \beta + \alpha_t$$

autoregression

$$\alpha_t = \phi\alpha_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2), \quad |\phi| < 1, \quad \theta = (\beta, \phi, \sigma^2)$$

likelihood

$$L(\theta; y_1, \ldots, y_n) = \int \left(\prod_{t=1}^{n} f(y_t \mid \alpha_t; \theta) \right) f(\boldsymbol{\alpha}; \theta) d\boldsymbol{\alpha}$$

$L_{approx}(\theta; y)$ via Laplace with some refinements

Davis & Yau, 2011

Some proposed solutions

- 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

Composite likelihood

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

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

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

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

- 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

Besag, 1975; Lindsay, 1988

Contemporary Mathematics
Volume **80**, 1988

1988a

COMPOSITE LIKELIHOOD METHODS

Bruce G. Lindsay[1]

ABSTRACT. Composite likelihood, sometimes called pseudolikelihood, is a likelihood type object formed by adding together individual component log likelihoods, each of which corresponds to a marginal or conditional event. A partial survey is made of the applications of this method, with emphasis made on methods for assessing, comparing, and improving efficiency. It is shown how structural information can be incorporated by conditioning on sufficient statistics. A new application based on rank likelihoods is introduced, and methods for assessing its information is given. Also, it is shown how to construct a stochastic Taylor series in an autonormal problem, with concomitant improvement in efficiency.

1. INTRODUCTION. In recent years there has been increased interest in a form of likelihood type estimation often called <u>pseudolikelihood</u>, first proposed by Besag [2]. We note that the name pseudolikelihood has been used in other contexts as well (e.g. [9]). With apologies to Besag, we will here use the term <u>composite likelihood</u> because it is descriptive of the method of construction we wish to consider.

We start with a parametric log likelihood $\mathfrak{L}(\beta;y)$, where y represents a vector valued random variable, and β an unknown p-dimensional real parameter. It is presumed that the problem has regularity and that, in particular, there exists a gradient $U(\beta)=\nabla\mathfrak{L}$, called the <u>efficient score function</u>, and Hessian $\nabla^2\mathfrak{L}$, where differentiation is with respect to the β vector. These are assumed to satisfy the usual relationship

---

1980 <u>Mathematics Subject Classification</u> (1985 <u>Revision</u>). 62F20.

[1]Supported by the National Science Foundation, DMS-8402735

This paper is in final form and no version of it will be submitted for publication elsewhere.

EXAMPLE 3C (*Composite rank likelihood*). Let $y_1,\ldots,y_n$ be the ordered values of an IID sample $x_1,\ldots,x_n$ from a continuous distribution $F_\beta(x)$. In the IID setting, constructing a composite likelihood using the components $x_i|x_{[i]}$ leads back to the usual likelihood. On the other hand, if we let $R_i(x)$ be the rank of observation x_i, and consider instead the likelihood of $R_i=r$ given $x_{[i]}$, we are lead to the component likelihood:

$$\mathfrak{L}_i(\beta) = \log\{F_\beta(y_{r+1})-F_\beta(y_{r-1})\}, \text{ where } r=R_i(x),$$

and the <u>composite</u> <u>rank</u> <u>likelihood</u>:

$$CL(\beta) = \Sigma \log\{F(y_{r+1})-F_\beta(y_{r-1})\}. \tag{3.8}$$

weighting components to increase efficiency of score equation

Wednesday session

Example: AR Poisson

Davis & Yau, 2011

- Likelihood

$$L(\theta; y_1, \ldots, y_n) = \int \left(\prod_{t=1}^{n} f(y_t \mid \alpha_t; \theta) \right) f(\boldsymbol{\alpha}; \theta) d\boldsymbol{\alpha}$$

- Composite likelihood

$$CL(\theta; y_1, \ldots, y_n) = \prod_{t=1}^{n-1} \int \int f(y_t \mid \alpha_t; \theta) f(y_{t+1} \mid \alpha_{t+1}; \theta) f(\alpha_t, \alpha_{t+1}; \theta) d\alpha_t d\alpha_{t+1}$$

- consecutive pairs
- Time-series asymptotic regime one vector y of increasing length
- Composite ML estimator still consistent, asymptotically normal, estimable asymptotic variance
- Efficient, relative to a Laplace-type approximation
- Surprises: AR(1), fully efficient; MA(1), poor; ARFIMA(0,d,0), ok

Variational methods

Titterington, 2006; Ormerod & Wand, 2010

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

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

- simple to calculate
  - $q(\theta) = \prod q_j(\theta_j)$
  - simple parametric family

- close to posterior: miminize Kullback-Leibler divergence between true posterior and approximation q

... variational methods

Titterington, 2006; Ormerod & Wand, 2010

- example GLMM:

$$\ell(\beta, \Sigma; y) = \log \int f(y \mid u; \beta) f(u; \Sigma) du$$

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

high-dimensional integral

- variational solution for some choice $q(u)$:

$$\ell(\beta, \Sigma; y) \geq \int q(u) \log\{ f(y, u; \beta, \Sigma) / q(u) \} du$$

- Simple choice of q : $N(\mu; \Lambda)$

variational parameters μ, Λ

Example: GLMM

Ormerod & Wand, 2012, *JCGS*

variational approx:

$$\ell(\beta,\Sigma) \ge \ell(\beta,\Sigma,\mu,\Lambda)$$
$$= \sum_{i=1}^m \left(y_i^{\mathrm{T}} X_i\beta - \tfrac{1}{2}\log|\Sigma|\right)$$
$$+ \sum_{i=1}^m E_{u\sim N(\mu_j,\Lambda_j)}\left(y_i^{\mathrm{T}} Z_i u - 1_i^{\mathrm{T}} b(X_i\beta + Z_i u) - \tfrac{1}{2}u^{\mathrm{T}}\Sigma^{-1}u - \log\{\phi_{\Lambda_j}(u-\mu_i)\}\right)$$

simplifies to k one-dim. integrals

- variational estimate:

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

- 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

Links to composite likelihood?

- 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. $\prod f(y_j; \theta)$
- S. Robin 2012 ”Some links between variational approximation and composite likelihoods?”
- Zhang & Schneider 2012 “A composite likelihood view for multi-label classification” JMLR V22
- Grosse 2015 “Scaling up natural gradient by sparsely factorizing the inverse Fisher matrix” ICML

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

Some proposed solutions

- 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

Indirect inference

- composite likelihood estimators are consistent

under conditions ...

- because $\log CL(\theta; y) = \sum_{i=1}^{n} \sum_{j<j'} \log f(y_j, y_{j'}; \theta)$
- derivative w.r.t. θ has expected value 0

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

- $\theta^* = \tilde{k}(\theta)$; 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 $\qquad$ Yi & R, 2010; Jiang & Turnbull, 2004

... indirect inference

Smith, 2008

- 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 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

- use simulated samples from model of interest
  to find the 'best' θ
- 'best' θ gives data that reproduces $\hat{\theta}^*$

Shalizi, 2013

... indirect inference

Smith, 2008

- simulate samples $y_t^m, \quad m = 1, \ldots, M$ at some value θ

  from the model

- 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 $p = d$ simply invert the 'bridge function'; if $p > d$, e.g.

$$\arg\min_\theta \{\hat{\theta}^*(\theta) - \hat{\theta}\}^{\mathrm{T}} W \{\hat{\theta}^*(\theta) - \hat{\theta}\}$$

- estimates of θ are consistent, asymptotically normal, but not efficient

… non-convex optimization

Grosse, 2015

Approximate Bayesian Computation

Marin et al., 2010

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

Fearnhead & Prangle, 2011

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

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

Some proposed solutions

- 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

Laplace approximation

$$\ell(\theta; y) = \log \int f(y \mid u; \beta) g(u; \Sigma) db = \log \int \exp\{Q(u, y, \theta)\} db, \text{say}$$

$\theta = (\beta, \Sigma)$

$$\ell_{Lap}(\theta; y) = Q(\tilde{u}, y, \theta) - \frac{1}{2} \log |Q''(\tilde{u}, y, \theta)| + c$$

using Taylor series expansion of $Q(\cdot, y, \theta)$ about $\tilde{u}$

simplification of the Laplace approximation leads to PQL:

$$\ell_{PQL}(\theta, u; y) = \log f(y \mid u; \beta) - \frac{1}{2} u^{\mathrm{T}} \Sigma^{-1} u$$

Breslow & Clayton, 1993

to be jointly maximized over u and θ and parameters in Σ

PQL can be viewed as linearizing $E(y)$ and then using results for linear mixed models Molenberghs & Verbeke, 2006

Extensions of Laplace approximations

- expansions valid with $p = o(n^{1/3})$ Shun & McCullagh, 1995
- expansions for mixed linear models to higher order
  Raudenbush et al., 2000
- use REML for variance parameters Nelder & Lee, 1996
- integrated nested Laplace approximation Rue et al., 2009
  - model $f(y_i \mid \theta_i)$; prior $\pi(\theta \mid \vartheta)$ parameters and hyper-par
  - posterior $\pi(\theta, \vartheta \mid y) \propto \pi(\theta \mid \vartheta)\pi(\vartheta) \prod f(y_i \mid \theta_i)$
  - marginal posterior

$$\pi(\theta_i \mid y) = \int \underbrace{\pi(\theta_i \mid \vartheta, y)}_{\text{Laplace}} \underbrace{\pi(\vartheta \mid y)}_{\text{Laplace}} d\vartheta$$

Quasi-likelihood

- simplify the model
- $$E(y_i;\theta) = \mu_i(\theta); \qquad \mathrm{Var}(y_i;\theta) = \phi\nu_i(\theta)$$
- consistent with generalized linear models
- example: over-dispersed Poisson responses
- PQL uses this construction, but with random effects

  Molenberghs & Verbeke, Ch. 14
- why does it work?
- score equations are the same as for a 'real' likelihood

  hence unbiased
- derivative of score function equal to variance function

  special to GLMs

Some proposed solutions

- 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

References