Composite Likelihood Estimation

With application to spatial clustered data

Cristiano Varin

Wirtschaftsuniversität Wien
April 29, 2016

Credits

CV, Nancy M Reid and David Firth (2011). An overview of composite likelihood methods. *Statistica Sinica*

Intractable likelihoods

Likelihoods often difficult to evaluate or specify in 'modern' (?) applications

Typical obstacles:

- large dense covariance matrices
- high-dimensional integrals
- normalization constants
- nuisance components
- ...

For example, models with unobservables

$$L(\theta; y) = \int f(y|u;\theta) f(u;\theta) \mathrm{d}u$$

Hard when the integral is high-dimensional like in spatial-temporal statistics

HOME | WHAT IS I-LIKE? | WHO IS I-LIKE? | EVENTS | PUBLICATIONS

i-like.org.uk

Intractable Likelihood

New Challenges from Modern Applications

A £2.4M EPSRC programme grant, **i-like**, aims to tackle some of the most important statistical challenges that arise across many modern day applications. It is led by Gareth Roberts (Warwick), and involves Christophe Andrieu (Bristol), Paul Fearnhead (Lancaster), David Firth (Warwick) and Chris Holmes (Oxford). See What is i-like? for more details.

NEWS

Workshop (22nd-24th June 2016). This year the i-like annual workshop will be held at the Lancaster University.
Details available here.

OLD NEWS

EPSRC

What are composite likelihoods?

Suppose intractable likelihood but low-dimensional distributions readily computed

Solution: combine low-dimensional terms to construct a pseudolikelihood

General setup:

- collection of marginal or conditional events $\{A_1, \ldots, A_K\}$
- associated component likelihoods $L_k(\theta; y) \propto f(y \in A_k; \theta)$

A composite likelihood is the weighted product

$$CL(\theta; y) = \prod_{k=1}^{K} L_k(\theta; y)^{w_k}$$

for some weights $w_k \geqslant 0$

Major credit. . .

Bruce G Lindsay (1988). Composite likelihood methods. *Contemporary Mathematics*

Marginal or conditional?

Marginal:

- independence likelihood $CL(\theta) = \prod_i f(y_i; \theta)$
- pairwise likelihood $CL(\theta) = \prod_i \prod_j f(y_i, y_j; \theta)$
- tripletwise $CL(\theta) = \prod_i \prod_j \prod_k f(y_i, y_j, y_k; \theta)$
- blockwise ...

Conditional:

- Besag pseudolikelihood
  $CL(\theta) = \prod_i f(y_i | \text{neighbours of } y_i; \theta)$
- full conditionals $CL(\theta) = \prod_i f(y_i | y_{(i)}; \theta)$
- pairwise conditional $CL(\theta) = \prod_i \prod_j f(y_i | y_j; \theta)$

Integrals...

Spatial generalized linear model:

$$E(Y_i|u_i) = g(x_i^\top \beta + u_i)$$

where u_i realization of a Gaussian random field

Likelihood function:

$$L(\theta; y) = \int_{\mathbb{R}^n} f(u_1, \ldots, u_n; \theta) \prod_{i=1}^{n} f(y_i|u_i; \theta) \mathrm{d}u_i$$

where $f(u_1, \ldots, u_n; \theta)$ is density of multivariate normal with dense covariance matrix

Pairwise likelihood:

$$PL(\theta; y) = \prod_i \prod_j \left\{ \int_{\mathbb{R}^2} f(u_i, u_j; \theta) f(y_i|u_i; \theta) f(y_j|u_j; \theta) \mathrm{d}u_i \mathrm{d}u_j \right\}^{w_{ij}}$$

Names...

Many names for just the same thing:

- composite likelihood
- pseudolikelihood
- quasi-likelihood
- limited information method
- approximate likelihood
- split-data likelihood
- ...

Comments:

- pseudo- and approximate likelihood too unspecific
- quasi-likelihood could be confused with the popular method for generalized linear models

Terminology

Log composite likelihood

$$cl(\theta) = \log CL(\theta)$$

Composite score

$$u_{cl}(\theta) = \partial cl(\theta)/\partial\theta$$

Maximum composite likelihood estimator

$$u_{cl}(\hat{\theta}_{cl}) = 0$$

Variability matrix

$$k(\theta) = \mathrm{Var}\{u_{cl}(\theta; Y)\}$$

Sensitivity matrix (Fisher information)

$$i(\theta) = \mathrm{E}\{-\partial u_{cl}(\theta; Y)/\partial\theta\}$$

Godambe information (sandwich information)

$$g(\theta) = i(\theta)k(\theta)^{-1}i(\theta)$$

Why it works?

Two arguments

First argument: The composite score function

$$u_{cl}(\theta) = \sum_k w_k \frac{\partial}{\partial \theta} \log L_k(\theta; y)$$

is a linear combination of 'valid' likelihood score functions

Unbiased under usually regularity conditions on each likelihood component

Asymptotic theory derived from standard estimating equations theory

Why it works? (cont'd)

Second argument: $\hat{\theta}_{\text{CL}}$ converges to the minimizer of the composite Kullback-Leibler divergence

$$CKL(\theta) = \sum_k w_k \mathrm{E}_h \left[\log \left\{ \frac{h(y \in A_k)}{f(y \in A_k; \theta)} \right\} \right]$$

where $h(\cdot)$ is 'density' of the 'true' model

For example, the maximum pairwise likelihood estimator converges to the minimizer of

$$CKL(\theta) = \sum_{(i,j)} w_{(i,j)} \int \log \left\{ \frac{h(y_i, y_j)}{f(y_i, y_j; \theta)} \right\} h(y_i, y_j) \, \mathrm{d}y_i \mathrm{d}y_j$$

Measure the distance from the true model only with bivariate aspects of the data

Apply directly the theory of misspecified likelihoods (White, 1982) with KL divergence replaced by CKL

Limit distribution

Y is an m-dimensional vector

Sample $y_1, \ldots, y_n$ from $f(y; \theta)$

Asymptotic consistency and normality for $n \to \infty$ and m fixed

$$\sqrt{n}(\hat{\theta}_{cl} - \theta) \sim N\{0, g(\theta)^{-1}\}$$

Sandwich-type asymptotic variance

$$g(\theta)^{-1} = i(\theta)^{-1} k(\theta) i(\theta)^{-1}$$

In the full likelihood case, we have $i(\theta) = k(\theta)$

More difficult if n fixed and $m \to \infty$, need assumptions on replication

For example, time series and spatial models require certain mixing properties

Significance functions

Composite likelihood versions of Wald and score statistics easily constructed

$$w_e(\theta) = (\hat\theta_{cl} - \theta)^\top g(\theta)(\hat\theta_{cl} - \theta) \xrightarrow{d} \chi^2_p \quad (\dim(\theta) = p)$$
$$w_u(\theta) = u_c(\theta)^\top g(\theta)^{-1} u_c(\theta) \xrightarrow{d} \chi^2_p$$

Composite likelihood ratio statistic with non-standard limit

$$w(\theta) = 2\{cl(\hat\theta_{cl}) - cl(\theta)\} \xrightarrow{d} \sum_{i=1}^{p} \lambda_i Z_i^2$$

with λ_i eigenvalues of $i(\theta)g(\theta)^{-1}$ and $Z_i \overset{iid}{\sim} N(0,1)$

Various proposals to 'calibrate' $w(\theta)$: Satterthwaite approx, rescaling, Saddlepoint (Pace et al., 2011)

Bayesian composite likelihoods

Composite posterior

$$\pi_c(\theta|y) = \frac{CL(\theta; y)\pi(\theta)}{\int CL(\theta; y)\pi(\theta)d\theta}$$

Overly precise inferences using directly the composite likelihood (Pauli et al., 2011; Ribatet et al., 2012)

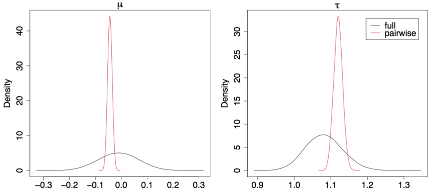


Figure 1: *Marginal full and pairwise posterior densities for the mean μ (left) and sill τ (right), derived from $n = 50$ realisations of a Gaussian process observed at $K = 20$ locations having an exponential covariance function with $\mu = 0$, $\tau = 1$ and $\omega = 3$.*

The curvature of CL needs to be adjusted... just the same problem of the composite likelihood ratio

Model selection

Model selection with the composite likelihood information criterion (Varin and Vidoni, 2005)

$$CLIC = -2cl(\hat{\theta}_{cl}) + 2\,trace\{i(\theta)^{-1}g(\theta)\}$$

Penalty $trace\{i(\theta)^{-1}g(\theta)\}$ accounts for the 'effective number of parameters'

Reduce to AIC when $i(\theta) = g(\theta)$

But reliable estimation of the model penalty often hard

Gong and Song (2011) derive BIC for composite likelihoods

Where are composite likelihood used?

Lots of application areas already, still growing rapidly

Popular application areas include

- genetics
- geostatistics
- correlated random effects (longitudinal data, time series, spatial models, network data)
- spatial extremes
- financial econometrics

Some references (already a bit outdated) in Varin, Reid and Firth (2011)

Efficiency?

Usually high efficiency when $n \to \infty$ and fixed m (longitudinal and clustered data)

Performance when $m \to \infty$ and n fixed (single long time series, spatial data) depends on the dependence structure

Some form of pseudo-replication is needed for acceptable efficiency when $m \to \infty$ and n fixed

Usually more efficient for discrete/categorical than continuous data

Carefull selection of likelihood components may improve efficiency

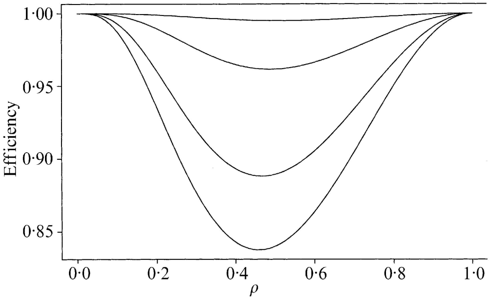
Symmetric normal

Cox and Reid (2004)

Efficiency of maximum pairwise likelihood for model

$$Y_i \overset{iid}{\sim} N_m(0, R) \quad Var(Y_{ir}) = 1 \quad Cor(Y_{ir}, Y_{is}) = \rho$$

(n independent vectors of size m)



Efficiency for fixed $m = 3, 5, 8, 10$

Truncated symmetric normal

Cox and Reid (2004)

Vectors of binary correlated variables generated truncating the symmetric normal model of the previous slide

Efficiency of maximum pairwise likelihood for $m = 10$:

ρ	.02	.05	.12	.20	.40	.50
ARE	.998	.995	.992	.968	.953	.968
ρ	.60	.70	.80	.90	.95	.98
ARE	.953	.903	.900	.874	.869	.850

Symmetric normal: large m fixed n

Cox and Reid (2004)

Symmetric normal

$$Var(\hat{\rho}_{pair}) = \frac{2}{n\, m(m-1)} \frac{(1-\rho^2)}{(1+\rho^2)^2} c(m^2, \rho^4)$$

$$\mathcal{O}(n^{-1}) \qquad \mathcal{O}(1)$$

$$n \to \infty \qquad m \to \infty$$

Truncated symmetric normal

$$Var(\hat{\rho}_{pair}) = \frac{1}{n} \frac{4\pi^2}{m^2} \frac{(1-\rho^2)}{(m-1)^2} c(m^4)$$

$$\mathcal{O}(n^{-1}) \qquad \mathcal{O}(1)$$

$$n \to \infty \qquad m \to \infty$$

not consistent if $m \to \infty$, n fixed!

Autoregressive model with additive noise

Varin and Vidoni (2009)

Autoregressive model with additive noise

$$Y_t = \beta + X_t + V_t, \quad V_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

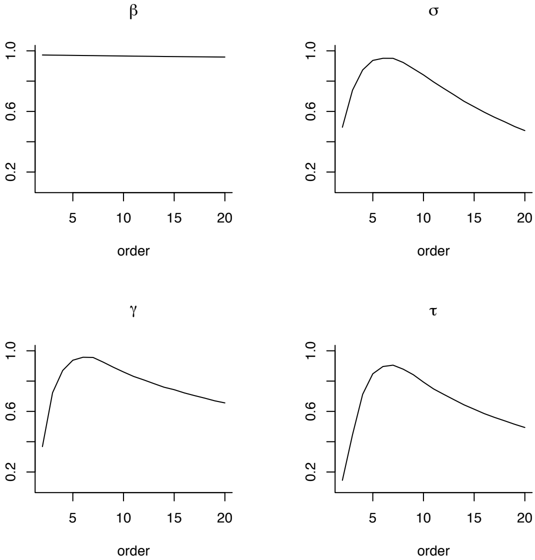
$$X_t = \gamma X_{t-1} + W_t, \quad W_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^2), \quad |\gamma| < 1$$

Pairwise likelihood of order d:

$$PL^{(d)}(\theta; y) = \prod_{r=d+1}^{n} \prod_{s=1}^{d} f(y_r, y_{r-s}; \theta)$$

In the special case of no observation noise ($\sigma^2 = 0$), $PL^{(1)}$ fully efficient. But $PL^{(d)}$ is increasingly inefficient as d increases.

What happens when there is observation noise?



Relative efficiency based on 1,000 simulated series of length 500 with $\beta = 0.1, \sigma = 1.0, \gamma = 0.95, \tau = 0.55$

Open areas

Composite likelihood general framework for scalable likelihood-type inference in complex models?

Perhaps, but there are several open questions to address first:

- choice of likelihood components
- choice of weights
- robustness
- reliable estimation of the variability matrix $k(\theta)$
- software implementation

Thanks for listening!