Composite Likelihoods
An Overview with Emphasis on Biostatistics

Cristiano Varin
Università Ca’ Foscari Venezia

with Nancy Reid, University of Toronto
Motivations

Setting: complex models where the ordinary likelihood function is difficult to evaluate or to specify.

Typical obstacles:
- inversion of large variance matrices
- approximation of intractable high dimensional integrals
- awkward normalization constants
- combinatorial difficulties
- nuisance components of difficult specification

In many of these situations it is however possible to compute marginal or conditional densities for subsets of the data.

**Composite likelihood idea:** construct a pseudolikelihood by combining conditional or marginal densities.  
Lindsay (1988)
Example I: Spatial generalized linear mixed models

GLMMs frequently used in Biostatistics for
- modeling heterogeneity in longitudinal data
- accounting for serial and spatial correlation
- ...

Inference based on the marginal likelihood obtained by integrating out the random effects.

Computational difficulties when random effects are correlated.

As in spatial GLMMs: Diggle and Ribeiro (2007)
- one random effect $u_i$ for each observation $Y_i$
- random effects are spatially correlated
- marginal likelihood is a high-dimensional integral

\[
L(\theta) = \int_{\mathbb{R}^n} \prod_{i=1}^{n} f(y_i|u_i; \theta) f(u_1, \ldots, u_n; \theta) \, du_1 \cdots du_n
\]
Marginal likelihood inference in spatial GLMMs requires approximating high-dimensional integrals.

Simulations? MCMC, MCEM, . . . , but need $O(n^3)$ operations!

Diggle and Ribeiro (2007)

Alternatives? composite likelihood constructed from bivariate marginal densities, a.k.a. pairwise likelihood:

$$CL^{(d)}(\theta) = \prod_{\{(i,j) \in S_d\}} \int_{\mathbb{R}^2} f(y_i|u_i; \theta)f(y_j|u_j; \theta)f(u_i, u_j; \theta)du_idu_j$$


Comments:

- product of much more innocuous bivariate integrals
- accurate deterministic quadrature approximations available
- only pairs of obs. close enough to be informative are used: $S_d = \{(i,j) : \text{distance between locations of } Y_i \text{ and } Y_j < d\}$
- computational cost reduced to $O(n)$ operations!
Example II: Robust inference with missing data

Liang and Qin (2000)

Target: correct inference on regressor coefficients $\beta$ in the regression model $f(y_i|x_i; \beta)$.

In many situations inference cannot be based solely on $f(y_i|x_i; \beta)$.

An important example is missing data: $\delta_i = 1$ if $(y_i, x_i)$ observed, $\delta_i = 0$ otherwise.

Observed likelihood?

$$L_{obs}(\theta) = \prod_{i=1}^{n} f(y_i, x_i | \delta_i = 1)$$

$$= \prod_{i=1}^{n} \frac{\Pr(\delta_i = 1|y_i)f(y_i|x_i; \beta)f(x_i; \phi)}{\Pr(\delta_i = 1)}$$

... but $L_{obs}$ suffers from lack of robustness: how to specify nuisance quantities $f(x_i; \phi)$ and $\Pr(\delta_i = 1|y_i)$?
Conditional arguments allow to avoid the specification of the nuisance quantities: 

- step 1: remove $f(x_i; \phi)$ by conditioning on observed $x_i$

$$L_{\text{cond}}(\theta) = \frac{\Pr(\delta_i = 1|y_i)f(y_i|x_i; \beta)}{\int \Pr(\delta_i = 1|y_i)f(y_i|x_i; \beta) dy_i}$$

but this still requires $\Pr(\delta_i = 1|y_i)$...

- step 2: also $\Pr(\delta_i = 1|y_i)$ removed by further conditioning on the order statistic

$$L_{\text{order}}(\beta) = \frac{\prod_{i=1}^n f(y_i|x_i; \beta)}{\sum_{\text{perm.}} \prod_{i=1}^n f(y(i)|x_i; \beta)}$$

Denominator is a sum over all possible permutations.

This conditional likelihood is very computational demanding: $\mathcal{O}(n!)$ operations!
Kalbfleisch’s conditional likelihood attractive but computable only for small sample sizes

$$L_{\text{order}}(\beta) = \frac{\prod_{i=1}^{n} f(y_i|x_i; \beta)}{\sum_{\text{perm.}} \prod_{i=1}^{n} f(y(i)|x_i; \beta)}$$

Feasible alternative: apply the same conditional argument in a “pairwise fashion”.

Composite likelihood based on bivariate conditional densities

$$CL(\beta) = \prod_{i<j} \frac{f(y_i|x_i; \beta)f(y_j|x_j; \beta)}{f(y_i|x_i; \beta)f(y_j|x_j; \beta) + f(y_i|x_j; \beta)f(y_j|x_i; \beta)}$$

Liang and Qin (2000)

$CL(\beta)$ does not require specification of anything but the regression model $f(y_i|x_i; \beta)$ and with only $O(n^2)$ operations!
Composite likelihoods

Ingredients:

- an \( m \)-dimensional vector random variable \( \mathbf{Y} \) with probability function/density \( f(\mathbf{y}; \theta) \)
- a set of marginal or conditional events \( \{A_1, \ldots, A_K\} \) with associated “sub-likelihoods” \( L_k(\theta) \propto f(\mathbf{y} \in A_k; \theta) \)
- a set of weights \( \{w_1, \ldots, w_K\} \)

**Composite likelihood**: inference function derived by multiplying the collection of sub-likelihoods

\[
CL(\theta) = \prod_{k=1}^{K} L_k(\theta)^{w_k}
\]

*Divide et Impera*: split a difficult problem into small problems that are easier to handle.
Names...

Composite likelihoods are refereed to with different names

- pseudo-likelihoods
- quasi-likelihoods
- limited information methods
- approximate likelihoods
- split-data likelihoods
- partial maximum likelihoods
- ...

...but these names are often misleading...
Terminology

**Pseudo-likelihood** any function of parameter and data that behaves under “some aspect” as a likelihood.

**Composite likelihoods** are just one of many examples of pseudo-likelihoods.

**Quasi-likelihood** used with different meanings, most common are
- statistics: the Wedderburn’s quasi-likelihood and variants;
- econometrics: quasi-likelihoods are *misspecified* likelihoods. For an econometrician, composite likelihoods are quasi-likelihoods because they are misspecified likelihoods.

**Limited information methods** in psychometrics are estimation procedures based on low dimensional margins.

Maydeu-Olivares (2006)
Derived quantities

Composite log-likelihood

\[ c\ell(\theta; y) = \sum_{k=1}^{K} w_k \ell_k(\theta; y) \]

Composite score

\[ u(\theta; y) = \nabla_\theta c\ell(\theta; y) = \sum_{k=1}^{K} w_k \nabla_\theta \ell_k(\theta; y) \]

\[ \mathbb{E}\{u(\theta; Y)\} = 0 \]

linear combination of scores

Maximum composite likelihood estimator

\[ \hat{\theta}_{\text{CL}} = \arg \max c\ell(\theta; y) \]

\[ u(\hat{\theta}_{\text{CL}}; y) = 0 \]

Sensitivity matrix (Fisher information)

\[ H(\theta) = \mathbb{E}_\theta \{-\nabla_\theta u(\theta; Y)\} \]

Variability matrix

\[ J(\theta) = \text{var}_\theta \{u(\theta; Y)\} \]

Godambe information (or sandwich information)

\[ G(\theta) = H(\theta)J(\theta)^{-1}H(\theta) \]

\[ H(\theta) \neq J(\theta) \]

misspecification
Justifications for composite likelihood inference

- **Unbiased estimating equations:** the composite score \( u(\theta; y) \) is a linear combination of valid scores, hence \( \mathbb{E}\{u(\theta; Y)\} = 0 \)
- \( \hat{\theta}_{\text{CL}} \) converges to the value minimizing the composite Kullback-Leibler divergence

\[
\text{CKL}(\theta) = \sum_{k=1}^{K} w_k \int \log \left\{ \frac{g(y \in A_k)}{f(y \in A_k; \theta)} \right\} g(y \in A_k) dy
\]

where \( g(y) \) is the true model density/probability function.

CKL is linear combination of Kullback-Leibler divergences, e.g. pairwise likelihood

\[
\text{CKL}_{\text{pair}}(\theta) = \sum_{i<j}^{k} w_{(i,j)} \int \log \left\{ \frac{g(y_i, y_j)}{f(y_i, y_j; \theta)} \right\} g(y_i, y_j) dy_i dy_j
\]

KL for pair \((y_i, y_j)\)
Asymptotic theory

Sample of i.i.d. $\mathbf{y}_1, \ldots, \mathbf{y}_n$, with $\text{dim}(\mathbf{y}_i) = m$, inference for $n \to \infty$ with $m$ fixed

$$CL(\theta; \mathbf{y}) = \prod_{i=1}^{n} CL(\theta; \mathbf{y}_i) \quad \text{and} \quad c\ell(\theta; \mathbf{y}) = \sum_{i=1}^{n} c\ell(\theta; \mathbf{y}_i)$$

Under “standard” regularity conditions

$$\sqrt{n}(\hat{\theta}_{CL} - \theta) \sim \mathcal{N}_p \{ \mathbf{0}, \mathbf{G}(\theta)^{-1} \}$$

$\mathbf{G}(\theta)$ is the Godambe information in $\mathbf{y}_1$.

More difficult the case of $n$ fixed and $m$ increasing, as one single ($n = 1$) long time series or a spatial dataset.

In such cases, the asymptotic theory depends on the availability of internal replication to obtain a central limit result.

Davis and Yau (2011)
Test statistics

Test $H_0: \psi = \psi_0$ with $\theta = (\psi, \tau)$, $\dim(\psi) = q \leq p$

CL versions of Wald and score test statistics easy to construct, with usual $\chi^2_q$ distribution from

$$\sqrt{n}(\hat{\theta}_{CL} - \theta) \sim \mathcal{N}_p (0, \mathbf{G}(\theta)^{-1})$$

As in ordinary likelihood inference, it is more attractive to consider the composite likelihood ratio statistic

$$W = 2 \left\{ c\ell(\hat{\theta}_{CL}) - c\ell(\psi_0, \hat{\tau}_{CL}(\psi_0)) \right\} \overset{d}{\to} \sum_{i=1}^{q} \lambda_i Z_i^2 \text{ (under } H_0\text{)}$$

with $Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and $\lambda_i$ eigenvalues of $\left(\mathbf{H}^{\psi\psi}\right)^{-1}\mathbf{G}^{\psi\psi}$

Kent (1982)

How to calibrate $W$? bootstrap, Satttherwaites approximation, Saddlepoint approximation, direct adjustments. Pace et al. (2011)
Model selection

Akaike’s information criterion

\[ \text{AIC} = -2 c\ell(\hat{\theta}_{CL}; y) + 2 \dim(\theta) \]

Varin and Vidoni (2005)

Bayesian information criterion

\[ \text{BIC} = -2 c\ell(\hat{\theta}_{CL}; y) + \log n \dim(\theta) \]

Gao and Song (2011)

Here, \( \dim(\theta) \) is the effective number of parameters estimated as

\[ \dim(\theta) = \text{tr}\{H(\theta)G(\theta)^{-1}\} \]

(but \( \dim(\theta) \) often really awkward to estimate with precision...)

Overview of Composite Likelihood Methods 15/37
Taxonomy

Composite conditional likelihoods: pseudo-likelihoods constructed by composition of conditional densities.

Composite marginal likelihoods: pseudo-likelihoods constructed by composition of marginal densities.

Hybrid methods: combination of composite likelihoods and other estimating functions or estimating equations.
Composite conditional likelihoods: examples

Besag’s pseudolikelihood

\[ CL(\theta) = \prod_{r=1}^{m} f(y_r | \{y_s : y_s \text{ neighbour of } y_r\}; \theta) \]

Besag (1974)

Stratified case-control studies

\[ CL(\theta) = \prod_{r=1}^{m-1} \prod_{s=r+1}^{m} f(y_r | y_r + y_s; \theta) \]


Pairwise conditional likelihood

\[ CL(\theta) = \prod_{r=1}^{m} \prod_{s=1}^{m} f(y_r | y_s; \theta) \]

Lindsay et al. (2011)

Full conditionals

\[ CL(\theta) = \prod_{r=1}^{m} f(y_r | y_{(-r)}; \theta) \]
Composite marginal likelihoods: examples

Independence likelihood

\[ CL(\theta) = \prod_{r=1}^{m} f(y_r; \theta) \]

Chandler and Bate (2007)

Pairwise (marginal) likelihood

\[ CL(\theta) = \prod_{r=1}^{m-1} \prod_{s=r+1}^{m} f(y_r, y_s; \theta) \]

Chandler and Bate (2007)

Tripletwise likelihood, . . . , blockwise

Eidsvik et al. (2012)

Pairwise differences

\[ CL(\theta) = \prod_{r=1}^{m-1} \prod_{s=r+1}^{m} f(y_r - y_s; \theta) \]

Curriero and Lele (1999)

Optimal combination of independence and pairwise likelihood

Cox and Reid (2004)
Hybrid methods

Hybrid pairwise likelihood: Kuk (2007)
- optimal estimating equations for parameters in the mean as in Liang and Zeger’ generalized estimating equations
- parameters in the variance estimated by composite likelihoods

Joint composite estimating functions: Bai et al. (2012)
- designed for spatio-temporal models
- construct separate composite scores for spatial, temporal and cross spatial-temporal components
- form a weighted quadratic objective function as in Hansen’s generalized method of moments
Applications

Many recent applications, some examples:

- **Longitudinal data**: Vasdekis et al. (2012)
- **Incomplete data**: Yi et al. (2011), Molenberghs et al. (2011)
- **Spatial statistics**: Bevilacqua et al. (2012), Bai et al. (2012)
- **Multivariate survival data**: Nielsen and Parner (2010)
- **Time series**: Davis and Yau (2011)
- **Genetics**: many papers here! see review of Larribe and Fearnhead (2011)
- **Spatial extremes**: Padoan et al. (2010), Davison and Gholamrezaee (2011)
- **Finance**: Engle and Kelly (2012)
- **Econometrics**: Bhat et al. (2012)

...
Symmetric normal
Cox and Reid (2004)

Symmetric normal model

\[ Y_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}_m(0, R) \]

\[ \text{var}(Y_{ir}) = 1 \quad \text{corr}(Y_{ir}, Y_{is}) = \rho \]

Pairwise likelihood (all pairs)

\[ \text{avar}(\hat{\rho}_{\text{pair}}) = \frac{2}{n m(m - 1)} \frac{(1 - \rho)^2}{(1 + \rho^2)^2} c(m^2, \rho^4) \]

Full likelihood

\[ \text{avar}(\hat{\rho}_{\text{full}}) = \frac{2}{n m(m - 1)} \frac{\{1 + (m - 1)\rho\}^2(\rho)^2}{1 + (m - 1)\rho^2} \]
Symmetric normal

Cox and Reid (2004)

Efficiency: lines correspond to \( m = 3, 5, 8, 10 \)
Truncated symmetric normal

Cox and Reid (2004)

Binary responses

\[ Y_{ir} = 1 \{ Z_{ir} > 0 \} \]

\[ \text{var}(Z_{ir}) = 1 \]

\[ Z_i \overset{i.i.d.}{\sim} N_m(0, R) \]

\[ \text{corr}(Z_{ir}, Z_{is}) = \rho \]

Pairwise likelihood (all pairs)

\[ \text{avar}(\hat{\rho}_{\text{pair}}) = \frac{1}{n} \frac{4\pi^2}{m^2 (m-1)^2} \text{var}(T), \quad T = \sum_{s<r} (2y_r y_s - y_r - y_s) \]

Full likelihood: asymptotic variance to be numerically evaluated

\[
\begin{array}{ccccccc}
\rho & 0.02 & 0.05 & 0.12 & 0.20 & 0.40 & 0.50 \\
\text{ARE} & 0.998 & 0.995 & 0.992 & 0.968 & 0.953 & 0.968 \\
\rho & 0.60 & 0.70 & 0.80 & 0.90 & 0.95 & 0.98 \\
\text{ARE} & 0.953 & 0.903 & 0.900 & 0.874 & 0.869 & 0.850 \\
\end{array}
\]

Efficiency for \( m = 10 \)
Truncated symmetric normal

Cox and Reid (2004)
... but if $m \to \infty$ and $n$ fixed ...

Symmetric normal

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

$O(n^{-1})$ $O(1)$

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

Truncated normal

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

$O(n^{-1})$ $O(1)$

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

Consistency fails if $m \to \infty$, $n$ fixed!
Headache Severity Diaries
Pain severity diaries

Diaries compiled by patients are very useful to document disease progression.

Severity of pain easier to measure on ordinal scales: very strong, strong, moderate, etc.

For chronic and recurrent pain conditions, such as migraine and back pain, studying the symptom severity over a time period is crucial to detect common- and person-specific pain trigger conditions.

Frequency of measurements can be (very) high using electronic data collection methods.
Headache in Toronto

Varin and Czado (2010)

Pain severity diaries about headache collected by psychologist T. Kostecki-Dillon

Diaries compiled by 119 patients

Four daily ratings: morning, noon, afternoon, bedtime

Diaries compiled for periods of different lengths: from four days (16 obs) to 213 consecutive days (852 obs)

The outcome is the severity of headache measured on an ordinal scale with six levels.

Long list of covariates about

- personal information
- clinical information
- weather conditions (does weather influence headache?)
Headache diaries: Serial dependence in mixed models

Cumulative mixed probit model:

• ordinal response $Y_{it}$ ($i = 1, \ldots, 119, t = 1, \ldots, n_i$) is a censored continuous latent variable

$$Y_{it} = y_{it} \iff \alpha_{y_{it} - 1} < Z_{it} \leq \alpha_{y_{it}}, \quad y_{it} \in \{1, \ldots, 6\},$$

where $-\infty \equiv \alpha_0 < \alpha_1 < \ldots < \alpha_6 \equiv \infty$

• unobserved $Z_{it}$ follows the linear mixed model

$$Z_{it} = x_{it} \top \beta + U_i + \epsilon_{it}, \quad U_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \quad \epsilon_{it} \sim \mathcal{N}(0, 1)$$

Since patient-specific series are long, it makes sense to account for serial dependence.

For example, by assuming errors $\epsilon_{it}$ following an AR(1) process

$$\epsilon_{it} = \rho \epsilon_{it-1} + \sqrt{1 - \rho^2} \eta_{it}, \quad \eta_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$$
Likelihood is the product of 119 integrals

\[ L(\theta) = \prod_{i=1}^{119} \int_{\alpha y_{i1} - 1}^{\alpha y_{i1}} \cdots \int_{\alpha y_{in} - 1}^{\alpha y_{in}} \frac{p(z_{i1}, \ldots, z_{in}; \theta)}{\text{multivariate normal}} dz_{i1} \cdots dz_{in} \]

The integrals dimensions \( n_1, \ldots, n_{119} \) range from 16 to 852.

Maximum likelihood is very cumbersome!

Feasible strategy: pairwise likelihood of order \( d \) constructed from patient-specific observations far apart not more than \( d \) time units

\[ CL^{(d)}(\theta) = \prod_{i=1}^{119} \prod_{t=d+1}^{n_i} \prod_{j=1}^{d} p(y_{it}, y_{it-j}; \theta) \]

\[ = \prod_{i=1}^{119} \prod_{t=d+1}^{n_i} \prod_{j=1}^{d} \int_{\alpha y_{it} - 1}^{\alpha y_{it}} \int_{\alpha y_{it-j} - 1}^{\alpha y_{it-j}} p(z_{it}, z_{it-j}; \theta) dz_{it} dz_{it-j} \]

This involves only two-dimensional integrals.
Pairwise likelihood order

Many papers show that it is not a good idea to use all possible pairs with time series, longitudinal data, and spatial data.

For the mixed ordinal probit model with AR(1) correlation used with headache-data model, the efficiency grows up until some distance:

![Graph showing the relationship between sum log variances and distance (d)].
Headache diaries: results

Model fitted with covariates:

- university degree?
- usage of analgesics?
- change in atmospheric pressure from previous day (3 levels)
- relative humidity index (3 levels)
- windchill (4 levels)

Different autocorrelation parameters for subjects with analgesics intake and for those without.

Results:

- graduated subjects have less severe symptoms (?)
- people with analgesics intake suffer from more severe headaches (rather obvious)
- no effect of humidity and windchill
- moderate effect of change in atmospheric pressure
Open questions

Does it matter if there is not a multivariate distribution compatible with, *e.g.*, bivariate margins? what about when no join distribution available?

Why does it work well? (when does it not work?)

How to investigate robustness systematically? Xu and Reid (2011)

“Design” issues: which terms should be included in composite likelihoods? *i.e.* pairs or triplets? conditional or marginal densities? and how they have to be combined? Lindsay et al. (2011)

How to estimate the Godambe information with precision in time series and spatial applications? (the “meat” matrix of the sandwich is difficult to estimate well)

How to construct sensible weighted versions of composite likelihood to improve on its efficiency?
Answers?

University of Warwick, April 2008
Banff Center, April 2012

Photos from the two workshop on composite likelihood methods
(I attended both, but I am not in the photos...)
Main references for this talk:

Merci beaucoup de votre attention!