Composite likelihood methods

Nancy Reid

University of Georgia
April 17, 2009

with Cristiano Varin
Ca' Foscari University, Venice

Likelihood function

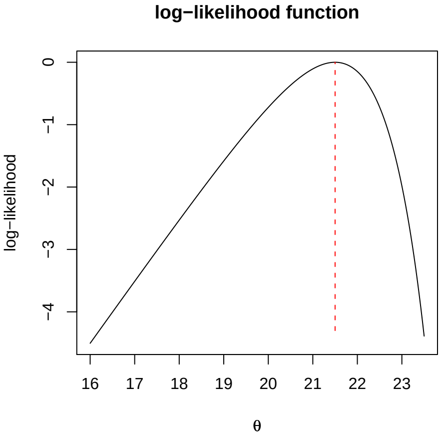
- Model: $Y \sim f(y;\theta)$
- Likelihood function: $L(\theta; y) = f(y;\theta) = \exp \ell(\theta; y)$

- Maximum likelihood estimate $\hat{\theta}$: $\sup_\theta L(\theta; y) = L(\hat{\theta}; y)$
- Score function: $U(\theta) = \ell'(\theta; y)$
- Fisher information:
  $\mathcal{I}(\theta) = E\{-\ell''(\theta; Y)\}, \quad j(\theta) = -\ell''(\theta; y)$

- Likelihood ratio: $w(\theta) = 2\{\ell(\hat{\theta}; y) - \ell(\theta; y)\}$

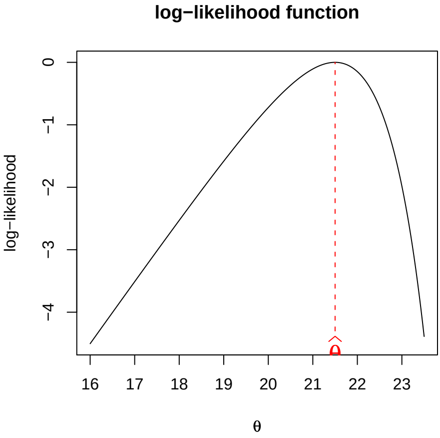
Likelihood inference

- Model: $Y \sim f(y; \theta)$
- Data: $y = (y_1, \ldots, y_n)$
- Likelihood function: $L(\theta; y) = \prod_{i=1}^{n} f(y_i; \theta)$



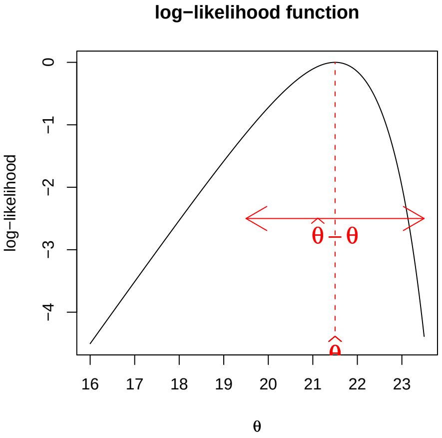
Likelihood inference

- Model: $Y \sim f(y;\theta)$
- Data: $y = (y_1, \ldots, y_n)$
- Likelihood function: $L(\theta; y) = \prod_{i=1}^{n} f(y_i;\theta)$



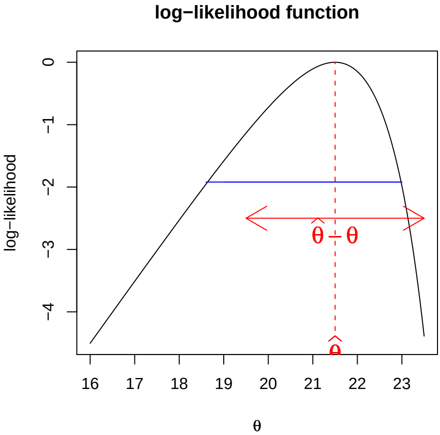
Likelihood inference

- Model: $Y \sim f(y;\theta)$
- Data: $y = (y_1, \ldots, y_n)$
- Likelihood function: $L(\theta; y) = \prod_{i=1}^{n} f(y_i;\theta)$



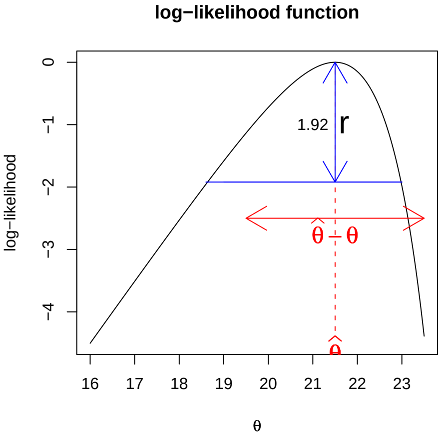
Likelihood inference

- Model: $Y \sim f(y; \theta)$
- Data: $y = (y_1, \ldots, y_n)$
- Likelihood function: $L(\theta; y) = \prod_{i=1}^{n} f(y_i; \theta)$



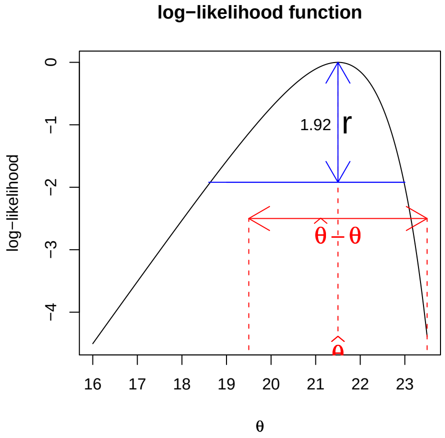
Likelihood inference

- Model: $Y \sim f(y;\theta)$
- Data: $y = (y_1, \ldots, y_n)$
- Likelihood function: $L(\theta; y) = \prod_{i=1}^n f(y_i;\theta)$



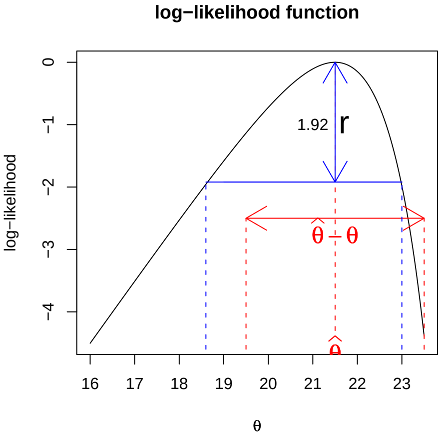
Likelihood inference

- Model: $Y \sim f(y; \theta)$
- Data: $y = (y_1, \ldots, y_n)$
- Likelihood function: $L(\theta; y) = \prod_{i=1}^{n} f(y_i; \theta)$



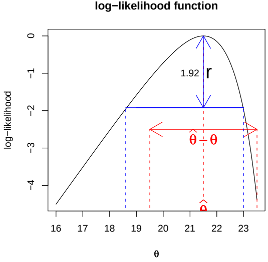
Likelihood inference

- Model: $Y \sim f(y; \theta)$
- Data: $y = (y_1, \dots, y_n)$
- Likelihood function: $L(\theta; y) = \prod_{i=1}^{n} f(y_i; \theta)$



Likelihood inference

$$\ell(\theta; y) = \Sigma \log f(y_i; \theta) \qquad j(\hat{\theta}) = -\ell''(\hat{\theta})/n \qquad \mathcal{I}(\theta) = E\{j(\theta)\}$$



stand'd mle: $\sqrt{n}(\hat{\theta} - \theta) \mathrel{\dot{\sim}} N\{0, j^{-1}(\hat{\theta})\}$

likelihood ratio: $w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \mathrel{\dot{\sim}} \chi^2_d$

standardized score: $U(\theta)/\sqrt{n} \mathrel{\dot{\sim}} N\{0, j(\hat{\theta})\}$

$$\frac{1}{\sqrt{n}} U(\theta) = \frac{1}{\sqrt{n}} \ell'(\theta) \xrightarrow{d} N\{0, \mathcal{I}(\theta)\}$$

Likelihood: advantages

- captures the sufficient statistics
- asymptotic theory for approximating distribution of derived quantities
- asymptotically fully efficient
- objective function for comparing models

Likelihood: drawbacks

- computation of derived quantities
- computation of the likelihood surface
- model misspecification
- high dimensional problems

Composite likelihood

- Model: $Y \sim f(y;\theta), \quad Y \in \mathcal{Y} \subset \mathbb{R}^p, \quad \theta \in \mathbb{R}^d$
- Set of events: $\{\mathcal{A}_k, k \in K\}$

- Composite Likelihood: (Lindsay, 1988)

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

- $L_k(\theta; y) = f(\{y_r \in \mathcal{A}_k\}; \theta)$ likelihood for an event
- $\{w_k, k \in K\}$ a set of weights

Examples

- Composite Conditional Likelihood: (Besag, 1974)

$$CCL(\theta; y) = \prod_{s \in \mathcal{S}} f_{s|s^c}(y_s \mid y_{s^c}),$$

  and variants by modifying events
- Composite Marginal Likelihood:

$$CML(\theta; y) = \prod_{s \in \mathcal{S}} f_s(y_s; \theta)^{w_s},$$

- Independence Likelihood: $\prod_{r=1}^{p} f_1(y_r; \theta)$
- Pairwise Likelihood: $\prod_{r=1}^{p-1} \prod_{s=r+1}^{p} f_2(y_r, y_s; \theta)$
- tripletwise likelihood, ...
- pairwise differences: $\prod_{r=1}^{p-1} \prod_{s=r+1}^{p} f(y_r - y_s; \theta)$
- and even mixtures of *CCL* and *CML*

Derived quantities

- log composite likelihood: $c\ell(\theta; y) = \log CL(\theta; y)$
- score function: $U(\theta; y) = \nabla_\theta c\ell(\theta; y) = \sum_{s \in \mathcal{S}} w_s U_s(\theta; y)$
- maximum composite likelihood est.: $\hat{\theta}_{CL} = \arg\sup c\ell(\theta; y)$

  $U(\hat{\theta}_{CL}) = 0$
- variability matrix: $J(\theta) = var_\theta\{U(\theta; Y)\}$
- sensitivity matrix: $H(\theta) = E_\theta\{-\nabla_\theta U(\theta; Y)\}$
- Godambe information (or sandwich information):

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

Inference

- Sample: $Y_1, \ldots, Y_n$, i.i.d., $CL(\theta; \underline{y}) = \prod_{i=1}^{n} CL(\theta; y_i)$
- $\sqrt{n}(\hat\theta_{CL} - \theta) \dot\sim N\{0, G^{-1}(\theta)\}$ $\qquad G(\theta) = H(\theta)J(\theta)^{-1}H(\theta)$
- $w(\theta) = 2\{c\ell(\hat\theta_{CL}) - c\ell(\theta)\} \dot\sim \sum_{a=1}^{d} \mu_a Z_a^2 \quad Z_a \sim N(0,1)$
- $\mu_1, \ldots, \mu_d$ eigenvalues of $J(\theta)H(\theta)^{-1}$
- $w(\psi) = 2\{c\ell(\hat\theta_{CL}) - c\ell(\tilde\theta_\psi)\} \dot\sim \sum_{a=1}^{d_0} \mu_a Z_a^2$
- constrained estimator: $\tilde\theta_\psi = \sup_{\theta=\theta(\psi)} c\ell(\theta; y)$
- $\mu_1, \ldots, \mu_{d_0}$ eigenvalues of $(H^{\psi\psi})^{-1} G^{\psi\psi}$
- Kent, 1982

Model selection

- Akaike's information criterion Varin and Vidoni, 2005

$$AIC = -2c\ell(\hat{\theta}_{CL}; y) - 2\,\mathrm{dim}(\theta)$$

- Bayesian information criterion Gao and Song, 2009

$$BIC = -2c\ell(\hat{\theta}_{CL}; y) - \log n\,\mathrm{dim}(\theta)$$

- effective number of parameters

$$\mathrm{dim}(\theta) = \mathrm{tr}\{H(\theta)G^{-1}(\theta)\}$$

- these criteria used for model averaging Hjort and Claeskens, 2008
- or for selection of tuning parameters Gao and Song, 2009

Example: symmetric normal

- $Y_i \sim N(0, R)$, $var(Y_{ir}) = 1$, corr $(Y_{ir}, Y_{is}) = \rho$
- compound bivariate normal densities to form pairwise likelihood

$$
\begin{aligned}
c\ell(\rho; y_1, \ldots, y_n) &= -\frac{np(p-1)}{4}\log(1-\rho^2) - \frac{p-1+\rho}{2(1-\rho^2)}SS_w \\
&\quad -\frac{(p-1)(1-\rho)}{2(1-\rho^2)}\frac{SS_b}{p}
\end{aligned}
$$

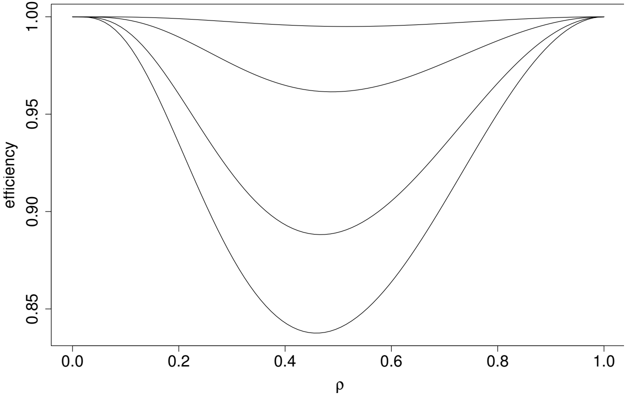
$$
SS_w = \sum_{i=1}^{n}\sum_{s=1}^{p}(y_{is} - \bar{y}_{i.})^2, \quad SS_b = \sum_{i=1}^{n} y_{i.}^2
$$

$$
\begin{aligned}
\ell(\rho; y_1, \ldots, y_n) &= -\frac{n(p-1)}{2}\log(1-\rho) - \frac{n}{2}\log\{1+(p-1)\rho\} \\
&\quad -\frac{1}{2(1-\rho)}SS_w - \frac{1}{2\{1+(p-1)\rho\}}\frac{SS_b}{p}
\end{aligned}
$$

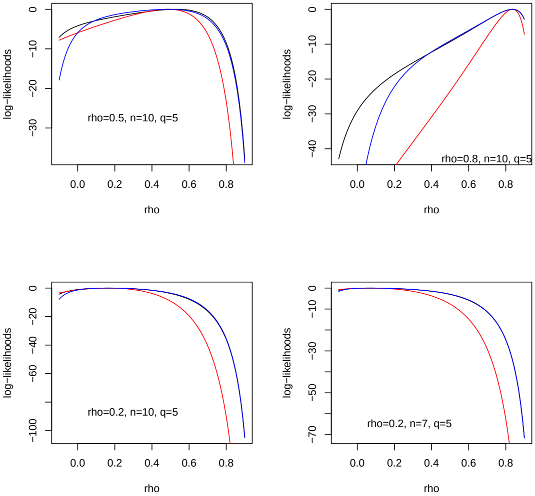
... symmetric normal

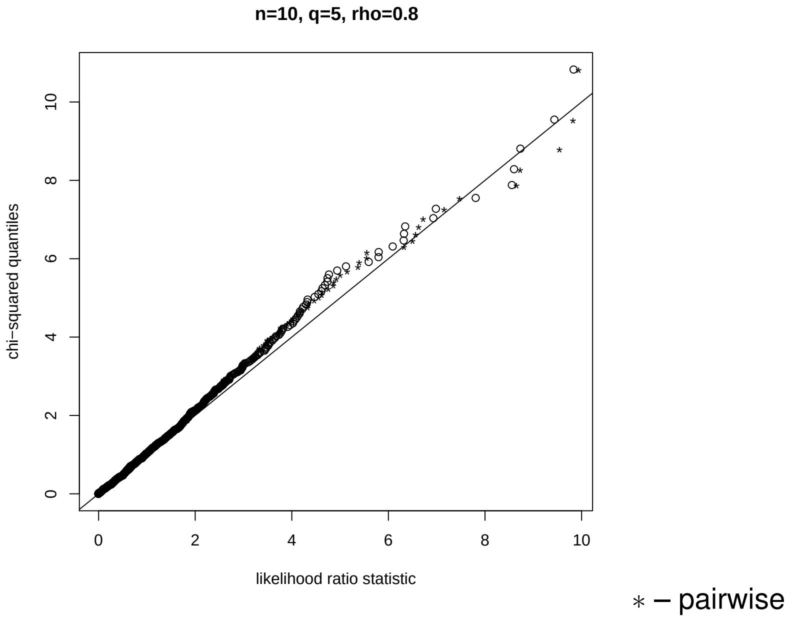
$$\frac{a.var(\hat{\rho}_{CL})}{a.var(\hat{\rho}\,)}, \quad p = 3, 5, 8, 10$$

(Cox & Reid, 2004)



Likelihood ratio test




n=10, q=5, rho=0.8
chi−squared quantiles
likelihood ratio statistic
0
2
4
6
8
10
0
2
4
6
8
10
∗ − pairwise

... symmetric normal +

- $Y_i \sim N(\mu \underline{1}, \sigma^2 R) \qquad R_{st} = \rho$
- $\hat{\mu} = \hat{\mu}_{CL}, \quad \hat{\sigma}^2 = \hat{\sigma}^2_{CL}, \quad \hat{\rho} = \hat{\rho}_{CL}$
- $G(\theta) = H(\theta) J(\theta)^{-1} H(\theta) = J(\theta)$
- pairwise likelihood is fully efficient
- also true for $Y_i \sim N(\mu, \Sigma)$
  (Mardia, Hughes, Taylor 2007; Jin 2009)

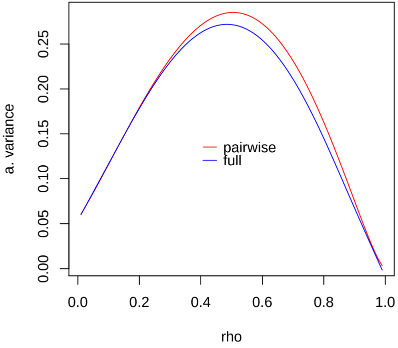
Example: dichotomized MV Normal

$$Y_r = 1\{Z_r > 0\} \qquad Z \sim N(0, R) \qquad r = 1, \dots, p$$

$$\ell_2(\rho) = \sum_{i=1}^{n} \sum_{s<r} \{y_r y_s \log P(y_r = 1, y_s = 1) + y_r(1 - y_s) \log P_{10} + (1 - y_r) y_s \log P_{01} + (1 - y_r)(1 - y_s) \log P_{00}\}$$

$$\text{a.var}(\hat{\rho}_{CL}) = \frac{1}{n} \frac{4\pi^2}{p^2} \frac{(1 - \rho^2)}{(p - 1)^2} \text{var}(T) \qquad T = \sum_{s<r} (2y_r y_s - y_r - y_s)$$

$$\text{var}(T) = p^4(p_{1111} - 2p_{111} + 2p_{11} - p_{11}^2 + \frac{1}{4}) + p^3(-6p_{1111} \dots) + p^2(\dots) + p(\dots)$$



ρ	0.02	0.05	0.12	0.20	0.40	0.50
ARE	0.998	0.995	0.992	0.968	0.953	0.968
ρ	0.60	0.70	0.80	0.90	0.95	0.98
ARE	0.953	0.903	0.900	0.874	0.869	0.850

Example: multi-level probit model

- latent variable: $z_{ir} = x_{ir}'\beta + b_i + \epsilon_{ir}, \quad \epsilon_{ir} \sim N(0,1)$
- binary observations: $y_{ir} = 1(z_{ir} > 0); \quad r = 1, \dots m_i; i = 1, \dots n$
- probit model: $Pr(y_{ir} = 1 \mid b_i) = \Phi(x_{ir}'\beta + b_i); \quad b_i \sim N(0, \sigma_b^2)$

- likelihood

$$L(\beta, \sigma_b) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \prod_{r=1}^{m_i} \Phi(x_{ir}'\beta + b_i)^{y_{ir}} \{1 - \Phi(x_{ir}'\beta + b_i)\}^{1-y_{ir}} \phi(b_i, \sigma_b^2) db_i$$

- pairwise likelihood

$$CL(\beta, \sigma_b) = \prod_{i=1}^{n} \prod_{r<s} P_{11}^{y_{ir}y_{is}} P_{10}^{y_{ir}(1-y_{is})} P_{01}^{(1-y_{ir})y_{is}} P_{00}^{(1-y_{ir})(1-y_{is})}$$

- each $Pr(y_{ir} = j, y_{is} = k)$ evaluated using $\Phi_2(\cdot, \cdot; \rho_{irs})$

(Renard et al., 2004)

... multi-level probit (Renard et al. 2004)

- computational effort doesn't increase with the number of random effects
- pairwise likelihood numerically stable
- efficiency losses, relative to maximum likelihood, of about 20% for estimation of β
- somewhat larger for estimation of σ_b^2

... Example

D. Renard et al. / Computational Statistics & Data Analysis 44 (2004) 649–667 663

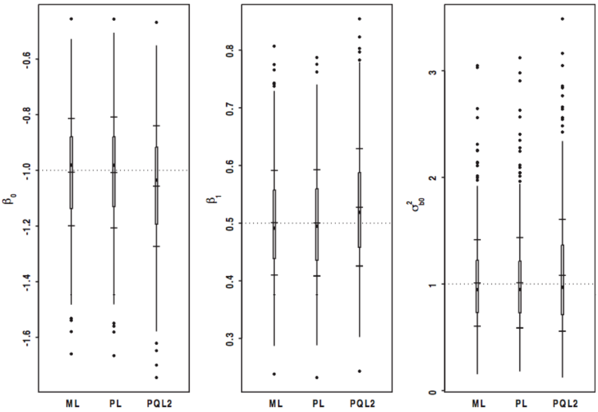


Fig. 5. Boxplots of ML, PL and PQL2 simulated parameter estimates under Model (10) with random intercept.

Markov chains Hjort and Varin, 2008

- comparison of likelihood

$$L(\theta; y) = \prod \mathrm{pr}(Y_r = y_r \mid Y_{r-1} = y_{r-1}; \theta)$$

- adjoining pairs CML

$$CML(\theta; y) = \prod \mathrm{pr}(Y_r = y_r, Y_{r-1} = y_{r-1}; \theta)$$

- composite conditional likelihood (= Besag's PL)

$$CCL(\theta; y) = \prod \mathrm{pr}(Y_r = y_r \mid \text{ neighbours }; \theta)$$

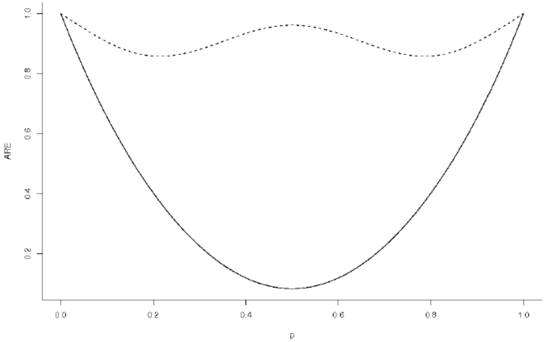
... Markov chain example

- Random walk with p states and two reflecting barriers
- Transition matrix

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 1-\rho & 0 & \rho & 0 & \dots & 0 \\ 0 & 1-\rho & 0 & \rho & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & 1 & 0 \end{pmatrix}$$

... Markov chain example

Reflecting barrier with five states: efficiency of pairwise likelihood (dashed line) and Besag's pseudolikelihood (solid line)



Continuous responses

- Multivariate Normal:
  $Y_i = (Y_{1i}, \dots, Y_{k_i i}) \sim N\{\beta_0 + \beta_1 x_i, \sigma^2 R_i(\alpha)\}$
  Zhao and Joe, 2005
- pairwise likelihood very efficient, but not $\equiv$ max. lik. ARE
- multivariate longitudinal data; correlated series of observations with random effects
  Fieuws and Verbeke,2006
- correlation of full likelihood and pairwise likelihood estimates of parameters near 1, relative efficiency also near 1 simulations
- pairwise likelihood based on differences within clusters, and connections to within and between block analysis
  Lele and Taper, 2002; Oakes and Ritz, 2000
- and several papers on survival data, often using copulas

CL2			
β_0	β_1	σ^2	ρ
0.998	0.997	1.000	0.913
0.996	0.995	1.000	0.889
0.995	0.996	0.999	0.876
1.000	0.999	1.000	0.884
0.960	0.968	0.987	0.967
0.974	0.970	0.993	0.964
0.978	0.969	0.992	0.928
0.986	0.977	0.993	0.903
0.942	0.958	0.961	0.957
0.944	0.949	0.961	0.952
0.949	0.945	0.966	0.922
0.964	0.939	0.966	0.898
0.924	0.966	0.934	0.943
0.926	0.947	0.937	0.940
0.943	0.932	0.949	0.925
0.982	0.913	0.976	0.919

Binary data

- $Y_r = 1\{Z_r > 0\}$, Z a latent normal r.v.
- generalizations to clustering, longitudinal data: Zhao and Joe 2005, Renard et al 2004
- random effects or multi-level models: Bellio and Varin, 2005; deLeon, 2004
- missing data: Parzen et al, 2007; Yi, Zeng and Cook, 2008
- YZC: not necessary to model the missing data mechanism, uses weighted pairwise likelihood, simulation results promising

... binary data

- questions re choice of weights with clustered data
- comparison of probit and logit
- not clear if marginal parameters and association parameters should be estimated separately
- mixed discrete and continuous data: deLeon and Carriere, 2006; Molenberghs and Verbeke, 2005
- Hybrid pairwise likelihood: GEE for marginal parameters and pairwise likelihood for association parameters: Kuk, 2007

Relation to Generalized Estimating Equations

- GEE specifies mean and variance, but not full model
- GEE is fully efficient in multivariate normal model with nonzero correlations
- pseudo likelihood is fully efficient in a specific multivariate binary model, with a particular dependence model ($\rho_{ir} \neq 0$, $\rho_{irs} \ldots$all zero)
- pseudo likelihood can fail badly, but most comparisons for clustered data are promising
- pseudo likelihood may be more robust to outliers than GEE
- (Qu and Song, 2004) discuss robustness of quadratic inference function
- pseudo-likelihoods are often easier to maximize
- example: network tomography (Liang and Yu, 2003)

Time series

- going back to proposal by Azzalini, 1983
- $n = 1$ of more interest, with long time series and possibly decaying correlations
- Markov chain models; using joint pairs
- e.g. in AR(1) fully efficient
- seems counterintuitive but seems to give good estimates
- state space models, population dynamics: Andrieu, 2008

And more...

- spatial data: multivariate normal, generalized linear models, CML based on differences, CCL and modifications, network tomography, data on a lattice, point processes
- image analysis: Nott and Ryden, 1999
- Rasch model, Bradley-Terry model, ...
- space-time data
- block-based likelihoods for geostatistics: Caragea and Smith, 2007
- gene mapping (linkage disequilibrium): Larribe and Lessard, 2008

- model selection using information criteria based on CL: Varin and Vidoni
- improvements of usual CL methods for specific models

Motivation for composite likelihood

- easier to compute:
  - binary data models with random effects, multi-level models (pairwise CML)
  - spatial data: ”near neighbours” CCL – Besag, 1974; Stein, Chi, Welty, 2004
  - sparse networks: Liang and Yu (2003)
  - long sequences (large p) in genetics: Fearnhead, 2003; Song, 2007
- access to multivariate distributions:
  - survival data: Parner, 2001; Andersen, 2004, using bivariate copulas
  - multi-type responses, such as continuous/discrete, missing data, extreme values, Oakes and Ritz,2000; deLeon, 2005; deLeon and Carriere, 2007
- more robust: model marginal (mean/variance) and association (covariance) parameters only

Questions about modelling

- Does it matter if there is not a multivariate distribution compatible with, e.g., bivariate margins?
- Does theory of multivariate copulas help in understanding this?
- How do we ensure identifiability of parameters?
  – examples of trouble?
- Relationship to modelling via GEE?
- In what sense is it more robust?
- E.g. binary data using dichotomized MV Normal

Questions about inference

- Efficiency of composite likelihood estimator:
  - choice of weights: Lindsay, 1988; Kuk and Nott, 2000;
  - assessment by simulation or direct comparison of a. var: Maydeu-Olivares and Joe, 2005
  - comparing two-stage to full pairwise estimation methods: Zhao and Joe, 2005; Kuk, 2007
  - ...

- Example: multivariate normal:
  - $Y \sim N(\underline{\mu}, \Sigma)$: pairwise likelihood estimates $\equiv$ mles
  - $Y \sim N(\mu\underline{1}, \sigma^2 R), R_{ij} = \rho$: pairwise likelihood est. $\equiv$ mles
  - $Y \sim N(\mu\underline{1}, R)$: loss of efficiency (although small)
- ? Why is CL so efficient (seemingly) ?

Questions about inference

- When Is CML (marginal) preferred to CCL (conditional) ? (always?)
- asymptotic theory: is composite likelihood ratio test preferable to Wald-type test?
- estimation of Godambe information: jackknife, bootstrap, empirical estimates
- estimation of eigenvalues of $(H^{\psi\psi})^{-1}G^{\psi\psi}$
- approximation of distribution of $w(\psi) \dot\sim \sum \mu_a Z_a^2$
  - Satterthwaite type? ($f\chi^2_d$): Geys et al, 1999
  - saddlepoint approximation?: Kuonen, 2004
  - bootstrap?
- large p, small n asymptotics: time series, genetics

$p \longrightarrow \infty$

- single long time series
- spatial models (p indexes spatial sites)
- usually assume decaying correlations, so p can play the role of n
- population genetics: estimation of the population recombination rate
- data is long sequence of alleles
- likelihood for each pair of segregating sites estimated by simulation
- pairwise likelihood formed by combining these
- Fearnhead & Donnelly, 2001; McVean et al., 2002; Fearnhead, 2003; Hudson, 2001

... $p \to \infty$

symmetric normal

$$\text{a.var}(\hat{\rho}_{CL}) = \frac{2}{np(p-1)} \frac{(1-\rho)^2}{(1+\rho^2)^2} c(p^2, \rho^4)$$

$$O(\tfrac{1}{n}) \qquad\qquad O(1)$$
$$n \longrightarrow \infty \qquad\qquad p \longrightarrow \infty$$

dichotomized mv normal:

$$\text{a.var}(\hat{\rho}_{CL}) = \frac{1}{n} \frac{4\pi^2}{p^2} \frac{(1-\rho^2)}{(p-1)^2} \text{var}(T)$$

$$\text{var}(T) = p^4(p_{1111} - 2p_{111} + 2p_{11} - p_{11}^2 + \frac{1}{4}) + p^3(-6p_{1111}\dots) + p^2(\dots) + p(\dots)$$

not consistent if $p \to \infty, n$ fixed

Questions

- is PL useful for modelling when no joint distribution is available (and may not exist?)
- e.g. extreme values, survival data (Parner, 2001)
- asymptotic theory in n, p together
- likelihood ratio type tests immediately available; one advantage over GEE
- can we really think beyond means and covariances in multivariate settings?
- should inference for mean parameters be separated from inference for covariances
- how to investigate robustness systematically
- estimation of Godambe information
- why does it work well? (when does it not work?)

References

- Varin, C. (2008) On composite marginal likelihoods. *Adv. Stat. Anal.* **95**, 1–28 www.dst.unive.it/~sammy
- www.utstat.utoronto.ca/reid/
- Lindsay, B. (1988) *Contemp. Math.* **80** 221–240
- Besag, J. (1974) *JRSS B* **34** 192–236
- Renard, D., Molenberghs, G. and Geys, H. (2004) *Comp. Stat. Data Anal.* **44** 629–667
- Kent, J. (1982) *Biometrika* **69** 19–27
- Cox, D.R. and Reid, N. (2004) *Biometrika* **91** 729–737
- Molenberghs, G. and Verbeke, G. (2005) *Models for discrete longitudinal data.* Springer-Verlag. [Ch. 9]
- Hjort and Varin (2008) *Scand. J. Statistics* **35**, 64–82
- Joe and Lee (2009) *J Multiv. Anal.* **100** 670–685

.. References

- Firth, Reid and Varin (2010?). An overview of composite likelihood methods. In preparation.
- Special issue of Statistica Sinica (editors Lindsay, Liang and Reid): http://www3.stat.sinica.edu.tw/statistica/