Composite likelihood methods

Nancy Reid

University of Warwick, April 15, 2008

Cristiano Varin

Grace Yun Yi, Zi Jin, Jean-François Plante

Some questions (and answers)

- Is there a drinks table at the poster session?
- I hope so, the Bayesians always have one and there are a lot of Bayesians here this week.
- Is it going to rain tomorrow?
- Quite possibly, but it could be worse, it could be snowing.
- Are you from Canada then?
- Eh?

Composite likelihood

- Model: $Y \sim f(y;\theta), \quad Y \in \mathcal{Y} \subset \mathbb{R}^p, \quad \theta \in \mathbb{R}^d$
- Composite Marginal Likelihood (CML):
  $CML(\theta; y) = \prod_{s\in\mathcal{S}} f_s(y_s;\theta), \quad \mathcal{S}$ is a set of indices
- Composite Conditional Likelihood (CCL):
  $CCL(\theta; y) = \prod_{s\in\mathcal{S}} f_{s|s^c}(y_s \mid y_{s^c}),$
- Independence Likelihood: $\prod_{r=1}^{p} f_1(y_r;\theta)$
- Pairwise Likelihood: $\prod_{r=1}^{p-1} \prod_{s=r+1}^{p} f_2(y_r, y_s;\theta)$
- Sample: $Y_1, \ldots, Y_n$, i.i.d., $CL(\theta; \underline{y}) = \prod_{i=1}^{n} CL(\theta; y_i)$

Composite likelihood

- Model: $Y \sim f(y;\theta), \quad Y \in \mathcal{Y} \subset \mathbb{R}^p, \quad \theta \in \mathbb{R}^d$
- Composite Marginal Likelihood (CML):
  $CML(\theta; y) = \prod_{s\in\mathcal{S}} f_s(y_s;\theta), \quad \mathcal{S}$ is a set of indices
- Composite Conditional Likelihood (CCL):
  $CCL(\theta; y) = \prod_{s\in\mathcal{S}} f_{s|s^c}(y_s \mid y_{s^c}),$
- Independence Likelihood: $\prod_{r=1}^{p} f_1(y_r;\theta)^{w_r}$
- Pairwise Likelihood: $\prod_{r=1}^{p-1} \prod_{s=r+1}^{p} f_2(y_r, y_s;\theta)^{w_r}$
- Sample: $Y_1, \ldots, Y_n$, i.i.d., $CL(\theta; \underline{y}) = \prod_{i=1}^{n} CL(\theta; y_i)$

Derived quantities

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

$$\begin{aligned} J(\theta) &= var_\theta\{U(\theta; Y)\} \\ H(\theta) &= E_\theta\{-\nabla_\theta U(\theta; Y)\} \end{aligned}$$

- Godambe information:

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

Inference

- $(\hat{\theta}_{CL} - \theta) \dot{\sim} N\{0, G^{-1}(\theta)\}$ $\quad G(\theta) = H(\theta)J(\theta)^{-1}H(\theta)$
- $w(\theta) = 2\{c\ell(\hat{\theta}_{CL}) - c\ell(\theta)\} \dot{\sim} \sum_{a=1}^{d} \mu_a Z_a^2 \quad Z_a \sim N(0,1)$
- $w(\psi) = 2\{c\ell(\hat{\theta}_{CL}) - c\ell(\tilde{\theta}_{\psi})\} \dot{\sim} \sum_{a=1}^{d_0} \mu_a Z_a^2$
- constrained estimator: $\tilde{\theta}_{\psi} = \sup_{\theta=\theta(\psi)} c\ell(\theta; y)$
- $\mu_1, \ldots, \mu_{d_0}$ eigenvalues of $(H^{\psi\psi})^{-1} G^{\psi\psi}$
- Sample: $Y_1, \ldots, Y_n, \quad n \to \infty$
- no nuisance parameters: eigenvalues of $H(\theta)^{-1}J(\theta)$
- θ scalar: scale factor J/H Kent, 1982

Example: symmetric normal

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

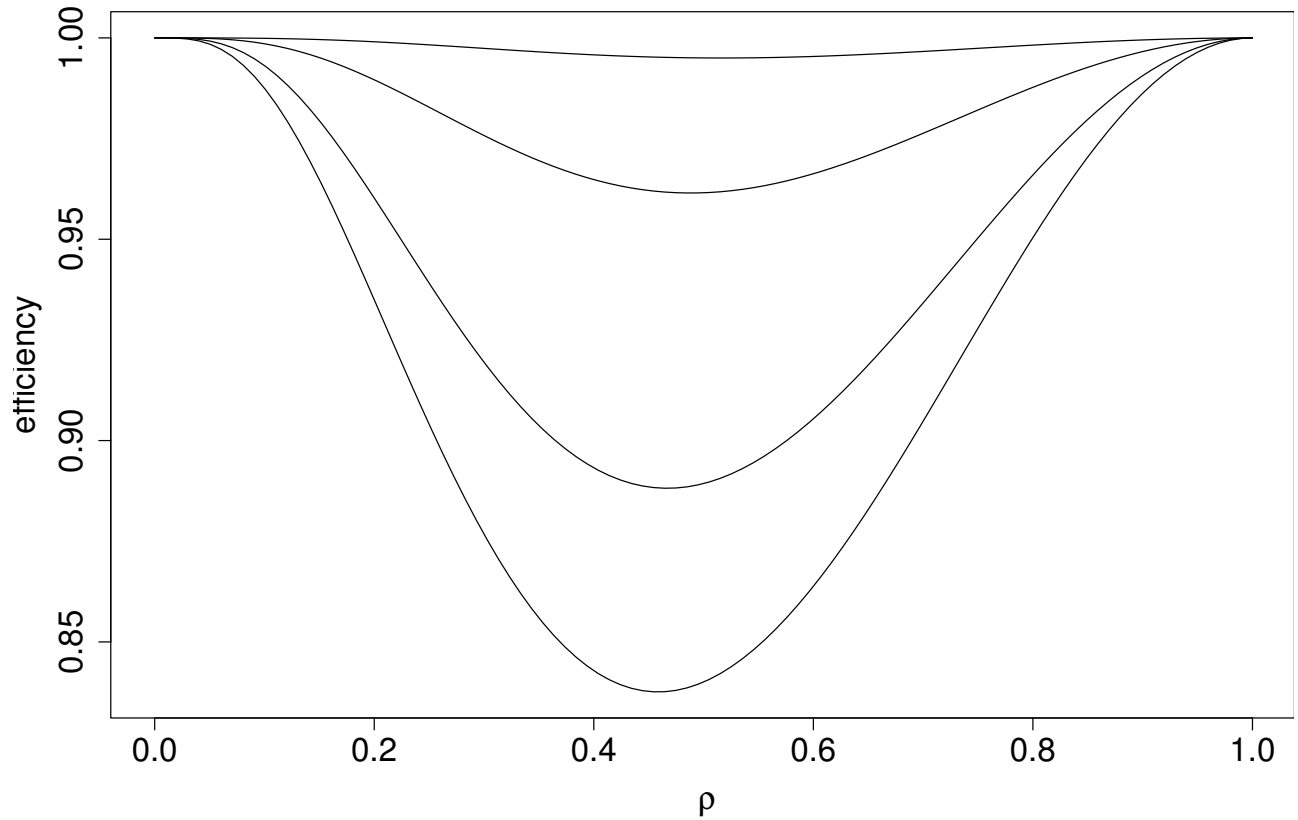
$$c\ell(\theta; y_1, \ldots, y_n) = -\frac{np(p-1)}{4}\log(1-\theta^2) - \frac{p-1+\theta}{2(1-\theta^2)}SS_w - \frac{(p-1)(1-\theta)}{2(1-\theta^2)}\frac{SS_b}{p}$$

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

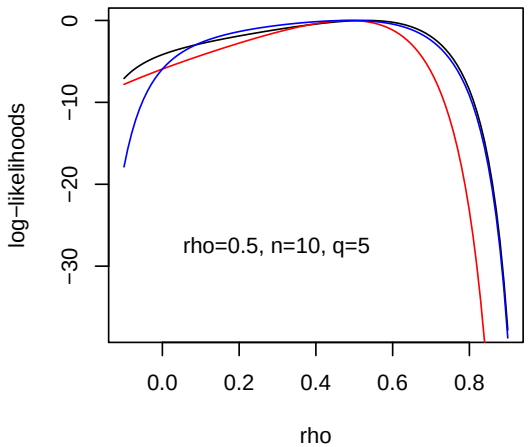
$$\ell(\theta; y_1, \ldots, y_n) = -\frac{n(p-1)}{2}\log(1-\theta) - \frac{n}{2}\log\{1+(p-1)\theta\} - \frac{1}{2(1-\theta)}SS_w - \frac{1}{2\{1+(p-1)\theta\}}\frac{SS_b}{p}$$

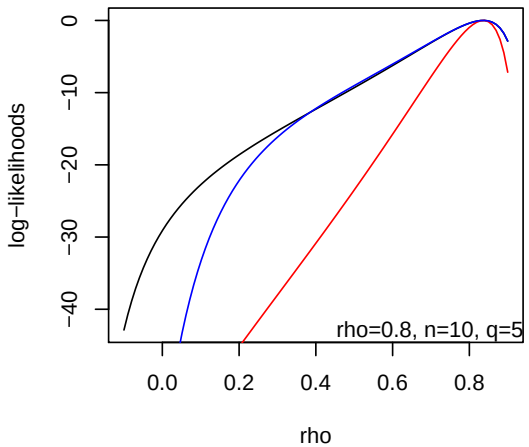
... symmetric normal

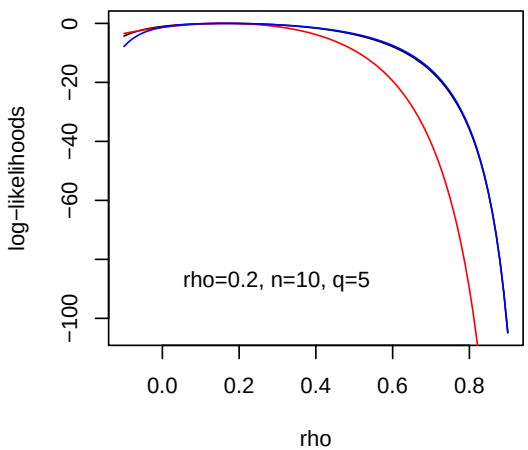
$$\frac{a.var(\hat{\theta}_{CL})}{a.var(\hat{\theta})}, \quad p = 3, 5, 8, 10$$

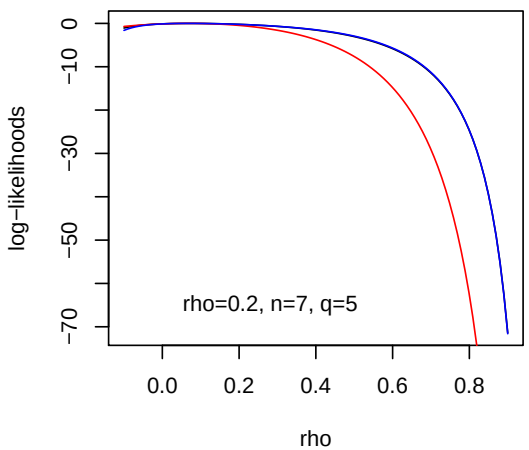


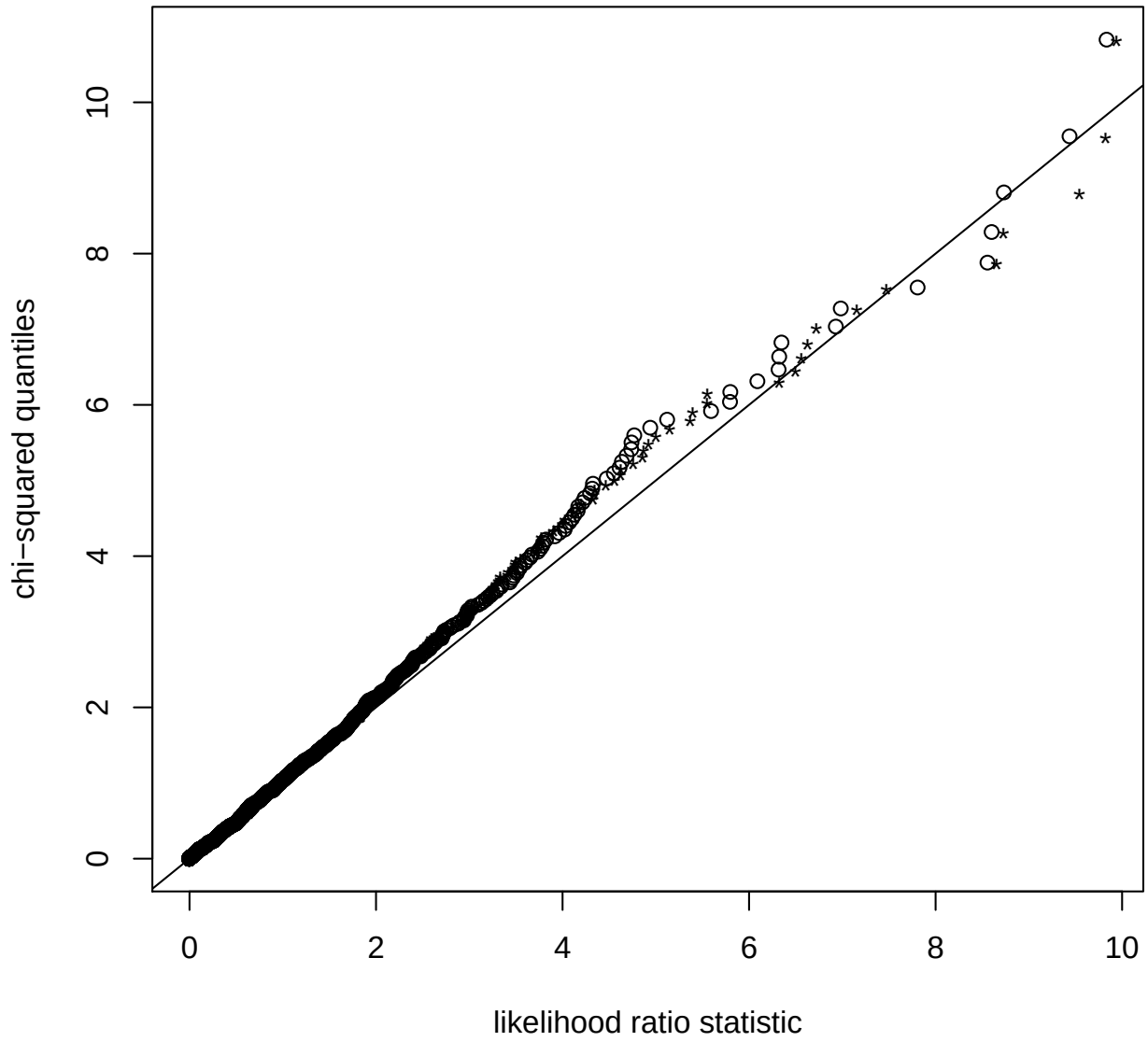
Likelihood ratio test










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

Motivation for composite likelihood

- easier to compute:
  - e.g. binary data models with random effects, multi-level models (pairwise CML)
  - e.g. spatial data (”near neighbours” CCL – Besag (1974); Stein, Chi, Welty (2004), Hellmund)
  - e.g. sparse networks (Liang and Yu, 2003), genetics (Fearnhead, Song), ...
- access to multivariate distributions:
  - e.g. survival data (Parner, 2001; Tibaldi, Barbosa, Molenberghs, 2004; Andersen, 2004), using bivariate copulas
  - e.g. multi-type responses, such as continuous/discrete, missing data, extreme values, Oakes and Ritz(2000), deLeon (2005), deL and Carriere (2007) Cattelan (Bradley-Terry model)
- 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? Hens
- Relationship to modelling via GEE?
- Connection to weighted likelihoods? (Zidek and Hu (1995), Plante)
- In what sense is it more robust? Liang
- E.g. binary data using dichotomized MV Normal

Questions about inference

- Efficiency of composite likelihood estimator: Lindsay, Joe, Fiocco, Bevilacqua, Kaimi, Zi, Oman
  - choice of weights: Lindsay, 1988; Kuk and Nott, 2000; Joe
  - assessment by simulation
  - comparing two-stage to full pairwise estimation methods (Zhao and Joe, 2005; Kuk, 2007)
  - ...

- Example: multivariate normal: Mardia, Hughes and Taylor (CCL), Jin (CML)
  - $Y \sim N(\underline{\mu}, \Sigma)$: pairwise likelihood estimates $\equiv$ mles
  - $Y \sim N(\mu \underline{1}, \sigma^2 R), R_{ij} = \theta$: pairwise likelihood est. $\equiv$ mles
  - $Y \sim N(\mu \underline{1}, R)$: loss of efficiency (although small)
  - dichotomized normal

- ? Can we *explain* efficiency?
  Hjort and Varin

Questions about inference

- When Is CML preferred to CCL? (always?)
  - CR 04: $c\ell(\theta) = \sum_{r,s} \log f_2(y_r, y_s; \theta) - a \sum_r \log f_1(y_r; \theta)$
- 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? $(f\chi_d^2)$ (Geys and Molen., Chandler
  - saddlepoint approximation? (Kuonen, 2004)
  - bootstrap?
- large p, small n asymptotics: not consistent;
  large p, medium n asymptotics?
  time series, genetics
- higher order asymptotic theory?

Continuous responses

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

- $Y = (Y_1, \ldots, Y_p)$, $Y_r = 1 \iff Z_r > 0$, Z a latent normal r.v. (CR 2004)
- 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

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 Hjort and Varin
- exact calculations for $AR(1)$, Jin
- Varin:

$$\prod_{t=m+1}^{n} \prod_{i=1}^{m} f_2(y_t, y_{t-i}; \theta)$$

- seems counterintuitive but seems to give good estimates
- state space models Andrieu, population dynamics

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
- genetics: Fearnhead, Song
- Rasch model, Bradley-Terry model, ...
- space-time data
- block-based likelihoods for geostatistics: Caragea and Smith, 2007

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