

Likelihood inference for complex data

Nancy Reid

Joint Statistical Meetings
August 4, 2009

with Cristiano Varin
Ca' Foscari University, Venice



Composite likelihood

- ▶ **Model:** $Y \sim f(y; \theta)$, $Y \in \mathcal{Y} \subset \mathbb{R}^p$, $\theta \in \mathbb{R}^d$
- ▶ **Set of events:** $\{\mathcal{A}_k, k \in K\}$
- ▶ **likelihood for an event:** $L_k(\theta; y) \propto f(\{y \in \mathcal{A}_k\}; \theta)$
- ▶ **Composite Likelihood:**

Lindsay, 1988

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

- ▶ $\{w_k, k \in K\}$ a set of weights

... composite likelihood

- ▶ Composite Conditional Likelihood = Pseudo-likelihood

Besag, 1974

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

- ▶ Composite Marginal Likelihood:

$$CML(\theta; y) = \prod_{s \in \mathcal{S}} f_s(y_s; \theta), \quad \text{subvectors}$$

- ▶ Independence Likelihood: $\prod_{r=1}^p f_1(y_r; \theta) \quad y = (y_1, \dots, y_p)$

- ▶ 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: $cl(\theta; y) = \log CL(\theta; y)$
- ▶ score function: $U(\theta; y) = \nabla_{\theta} cl(\theta; y) = \sum_{s \in \mathcal{S}} U_s(\theta; y)$
 $E\{U(\theta; Y)\} = 0$
- ▶ maximum composite likelihood est.: $\hat{\theta}_{CL} = \arg \sup_{\theta} cl(\theta; y)$
 $U(\hat{\theta}_{CL}) = 0$
- ▶ variability matrix: $J(\theta) = \text{var}_{\theta}\{U(\theta; Y)\}$
- ▶ sensitivity matrix: $H(\theta) = E_{\theta}\{-\nabla_{\theta} U(\theta; Y)\}$
- ▶ Godambe information (or sandwich information):

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

- ▶ $J \neq H$

misspecified model

Inference

- ▶ **Sample:** Y_1, \dots, Y_n $CL(\theta; y) = \prod_{i=1}^n CL(\theta; y_i)$
- ▶ $\sqrt{n}(\hat{\theta}_{CL} - \theta) \sim N\{0, G^{-1}(\theta)\}$ $G(\theta) = H(\theta)J(\theta)^{-1}H(\theta)$
- ▶ $w(\theta) = 2\{cl(\hat{\theta}_{CL}) - cl(\theta)\} \sim \sum_{a=1}^d \mu_a Z_a^2$ $Z_a \sim N(0, 1)$
- ▶ μ_1, \dots, μ_d eigenvalues of $J(\theta)H(\theta)^{-1}$
- ▶ $w(\psi) = 2\{cl(\hat{\theta}_{CL}) - cl(\tilde{\theta}_\psi)\} \sim \sum_{a=1}^{d_0} \mu_a Z_a^2$
- ▶ constrained estimator: $\tilde{\theta}_\psi = \arg \sup_{\theta=\theta(\psi)} cl(\theta; y)$
- ▶ μ_1, \dots, μ_{d_0} eigenvalues of $(H^{\psi\psi})^{-1}G^{\psi\psi}$
- ▶

Kent, 1982

Model selection

- ▶ Akaike's information criterion Varin and Vidoni, 2005

$$AIC = -2cl(\hat{\theta}_{CL}; y) - 2 \dim(\theta)$$

- ▶ Bayesian information criterion Gao and Song, 2009

$$BIC = -2cl(\hat{\theta}_{CL}; y) - \log n \dim(\theta)$$

- ▶ effective number of parameters

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

- ▶ model averaging Hjort and Claeskens, 2008
- ▶ selection of tuning parameters Gao and Song, 2009

Some recent applications

Longitudinal data, binary and continuous: random effects models

Molenberghs and Verbeke, 2005, Ch. 9; Zhao & Joe, 2005

Survival analysis: frailty models, copulas

Parner, 2001; Andersen, 2004; Fiocco et al., 2009

Multi-type responses: discrete and continuous; markers and event times

de Leon and Carriere, 2007; Fieuws et al., 2007

Finance: time-varying covariance models

Engle et al., 2009

Genetics/bioinformatics: large literature

Tamura et al., 2007; Li, 2008; Mardia et al., 2009

CCL for vonMises distribution: protein folding

Spatial data: geostatistics, spatial point processes

Stein, 2004; Caragea and Smith, 2008; Varin et al., 2005; ...

and more...

- ▶ image analysis Nott and Ryden, 1999
- ▶ genetics Fearnhead, 2008; Song, 2008
- ▶ gene mapping (linkage disequilibrium) Larribe and Lessard,
2008
- ▶ Rasch model, Bradley-Terry model, ...
- ▶ state space models, population dynamics: Andrieu, 2008

Example: symmetric normal

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

$$cl(\rho; y_1, \dots, y_n) = -\frac{np(p-1)}{4} \log(1-\rho^2) - \frac{\rho-1+\rho}{2(1-\rho^2)} SS_w - \frac{(\rho-1)(1-\rho)}{2(1-\rho^2)} \frac{SS_b}{\rho}$$

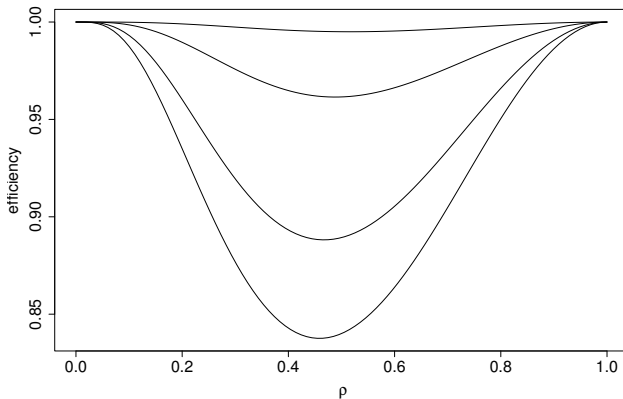
$$SS_w = \sum_{i=1}^n \sum_{s=1}^p (y_{is} - \bar{y}_i)^2, \quad SS_b = \sum_{i=1}^n y_i^2$$

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

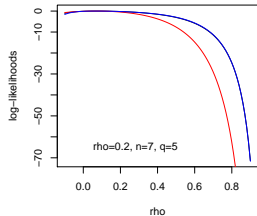
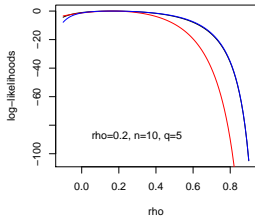
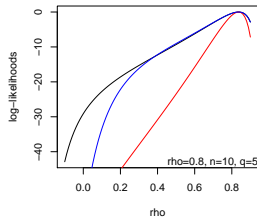
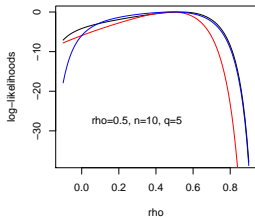
... symmetric normal

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

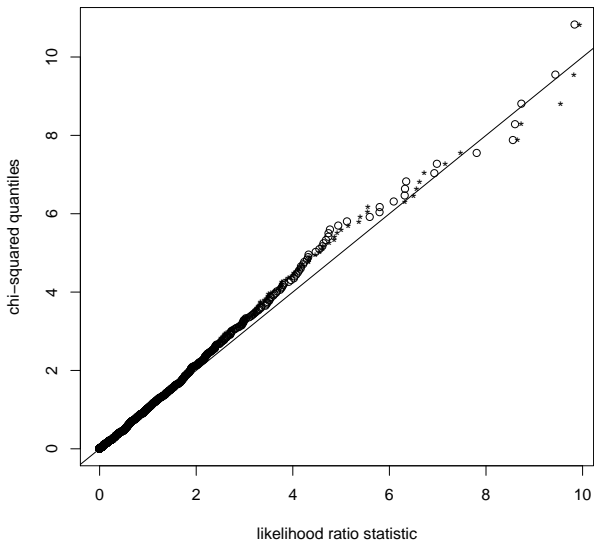
(Cox & Reid, 2004)



Likelihood ratio test



$n=10, q=5, \rho=0.8$



* – pairwise

... symmetric normal +

▶ $Y_i \sim N(\underline{\mu}, \sigma^2 R) \quad R_{st} = \rho$

▶ $\hat{\mu} = \hat{\mu}_{CL}, \quad \hat{\sigma}^2 = \hat{\sigma}_{CL}^2, \quad \hat{\rho} = \hat{\rho}_{CL}$

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

▶ pairwise likelihood is fully efficient

▶ also true for $Y_i \sim N(\mu, \Sigma)$

(Mardia, Hughes, Taylor 2007; Jin 2009)

Markov chains Hjort and Varin, 2008

- ▶ comparison of likelihood

$$L(\theta; y) = \prod_{r=2}^p \text{pr}(Y_r = y_r \mid Y_{r-1} = y_{r-1}; \theta)$$

- ▶ adjoining pairs CML

$$CML(\theta; y) = \prod_{r=1}^p \text{pr}(Y_r = y_r, Y_{r-1} = y_{r-1}; \theta)$$

- ▶ composite conditional likelihood (= Besag's PL)

$$CCL(\theta; y) = \prod_{r=2}^{p-1} \text{pr}(Y_r = y_r \mid \text{neighbours}; \theta)$$

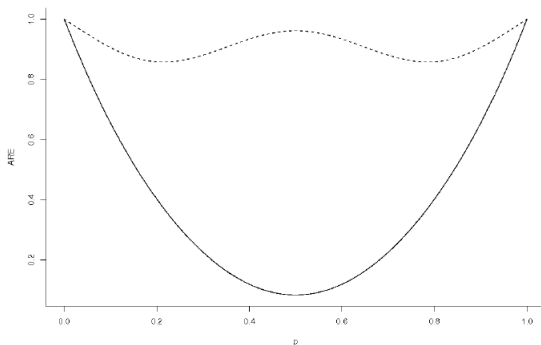
... Markov chain example

- ▶ Random walk with ρ states and two reflecting barriers
- ▶ Transition matrix

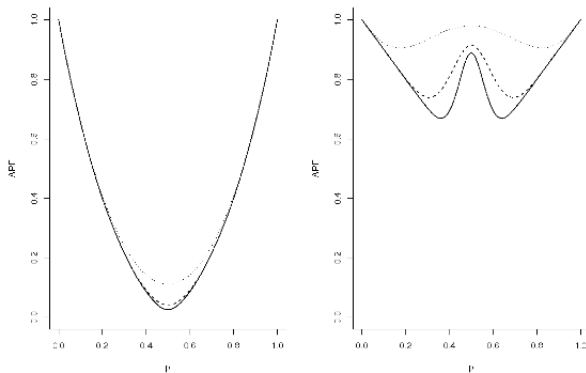
$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 1 - \rho & 0 & \rho & 0 & \dots & 0 \\ 0 & 1 - \rho & 0 & \rho & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & 1 & 0 \end{pmatrix}$$

... Markov chain example

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



... increasing the number of states



Effect on increasing the number of states: Besag's PL (left) and pairwise likelihood (right). The lines correspond to 5, 10, 15 states.

Aspects of efficiency

- ▶ $n \rightarrow \infty$, e.g. longitudinal data: pairwise likelihood estimates have high efficiency
- ▶ $p \rightarrow \infty$, fixed n , e.g. single long series, genetics, spatial data
- ▶ performance depends strongly on the dependence structure: need pseudo-replication
- ▶ e.g. not suitable for long memory processes
- ▶ fine for parameter-driven models in time series
- ▶ usually more efficiency loss for continuous than discrete/categorical responses Joe and Lee, 2009
- ▶ choosing weights in $CL = \prod f(y_s; \theta)^{w_s}$ to increase efficiency Zhao and Joe, 2005; Lindsay, 1988

Aspects of robustness

- ▶ model robustness
 - ▶ univariate and bivariate margins only for example
 - ▶ means, variances, association parameters
 - ▶ similar in flavour to generalized estimating equations GEE:
mean structure primary
- ▶ computational robustness
 - ▶ composite log-likelihood functions are **smoother** than log-likelihood functions
 - ▶ easier to maximize, easier to work with
 - ▶ especially in high dimension cases Liang and Yu, 2003
- ▶ adapting the EM algorithm: Song and Gao
- ▶ robust to missing data mechanisms: Yi, Zeng and Cook
- ▶ efficiency and robustness in binary data: Zi Jin, # 607
- ▶ access to multivariate distributions: extreme values, survival data, ...

Questions about inference

- ▶ When Is composite **marginal** likelihood preferred to **composite** conditional likelihood ? (always?)
- ▶ why is composite likelihood seemingly so efficient?
 - ▶ $Y \sim N(\underline{\mu}, \Sigma)$: pairwise likelihood estimates \equiv mles
 - ▶ $Y \sim N(\underline{\mu}_1, \sigma^2 R)$, $R_{ij} = \rho$: pairwise likelihood est. \equiv mles
 - ▶ $Y \sim N(\underline{\mu}_1, R)$: loss of efficiency (although small)
- ▶ asymptotic theory: is composite likelihood ratio test preferable to Wald-type test?
- ▶ estimation of Godambe information $J = \text{var}U(\theta)$
jackknife, bootstrap, empirical estimates
- ▶ approximation of distribution of $w(\psi) \sim \sum \mu_a Z_a^2$
- ▶ **large p , small n asymptotics: time series, genetics**

$$\dots p \rightarrow \infty$$

symmetric normal

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

$$\begin{matrix} O\left(\frac{1}{n}\right) \\ n \rightarrow \infty \end{matrix}$$

$$\begin{matrix} O(1) \\ p \rightarrow \infty \end{matrix}$$

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

Questions about modelling

- ▶ Is CL useful for modeling when no multivariate distribution exists that is compatible with margins?
- ▶ e.g. extreme values, survival data Parner, 2001
- ▶ Does theory of multivariate copulas help in understanding this?
- ▶ How do we ensure identifiability of parameters?
– examples of trouble?
- ▶ Relationship to modelling via GEE?
- ▶ how to investigate robustness systematically?
- ▶ how to make use of objective function
- ▶ can we really think beyond means and covariances in multivariate settings?

References

- ▶ Varin, C. (2008) *Adv. Stat. Anal.* **92**, 1–28.
www.dst.unive.it/~sammy
- ▶ Lindsay, B. (1988) *Contemp. Math.* **80** 221–240
- ▶ 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
- ▶ Firth, Reid and Varin (2010?). An overview of composite likelihood methods. In preparation.

Special issue of *Statistica Sinica*

<http://www3.stat.sinica.edu.tw/statistica/>