

Likelihood-based inference in complex models

Nancy Reid

York University

February 17, 2006

David Cox (Oxford), Grace Yun Yi (Waterloo), Jin Zi (Toronto)

Likelihood inference

- parametric model $Y \sim f(y; \theta)$ $y \in R^q$, $\theta \in R^p$
- data y_1, \dots, y_n
- likelihood $L(\theta; y_1, \dots, y_n) \propto \prod_{i=1}^n f(y_i; \theta)$
- log-likelihood $\ell(\theta) = \sum_{i=1}^n \log f(y_i; \theta) +$
- likelihood inference

$$(\hat{\theta} - \theta)^T \{-\ell''(\hat{\theta})\}(\hat{\theta} - \theta) \quad \sim \quad \chi_p^2$$

$$\ell'(\theta)^T \{-\ell''(\hat{\theta})\}^{-1} \ell'(\theta) \quad \sim \quad \chi_p^2$$

$$2\{\ell(\hat{\theta}) - \ell(\theta)\} \quad \sim \quad \chi_p^2$$

- asymptotic theory requires $\hat{\theta} \xrightarrow{p} \theta$ $\ell'(\hat{\theta}) = 0$, $E\ell'(\theta) = 0$
- a central limit theorem for $\ell'(\theta)$
- $E(\ell' \ell'^T) = E(-\ell'')$ and they are $O(n)$
- some further regularity conditions on the model...

Some difficulties

- ‘nonregular’ models: endpoint parameters, changepoint problems, strong dependence, not enough aggregation
- inference for subparameters: the approximations are too crude, and need to be adjusted for nuisance parameters
- data has complex structure: spatial data, population genetics, longitudinal data, clustered data, ...
- plausible models exist, but difficult to evaluate
- plausible models exist but are not reliable
- simpler, and/or more ‘robust’ modelling preferred:
 - mean/variance specifications as in generalized linear models and quasi-likelihood
 - generalized estimating equations (GEE) with ‘working’ covariance structure

Pseudo-likelihood

(Cox and R, 2004)

- single observation $y = (y_1, \dots, y_q) \sim f(y; \theta)$
- combine marginals $f_s(y_s)$ and bivariate marginals $f_{rs}(y_r, y_s)$

$$\begin{aligned} \ell_2(\theta; y) &= \sum_{r < s} \log f_{rs}(y_r, y_s) - aq \sum_s \log f_s(y_s) \\ &= \sum_{r < s} \log f_{rs}(y_r, y_s) - aq \ell_1(\theta; y) \end{aligned}$$

- n independent observations

$$\ell_2(\theta; y_1, \dots, y_n) = \sum_{i=1}^n \ell_2(\theta; y_i)$$

- score function $U_2(\theta; y_1, \dots, y_n) = \partial \ell_2(\theta; y_1, \dots, y_n) / \partial \theta$

Estimation of θ from U_2

- $U_2(\tilde{\theta}) = 0$
- $U_2(\tilde{\theta}) = U_2(\theta) + (\tilde{\theta} - \theta)^T \frac{\partial U_2(\tilde{\theta})}{\partial \theta} + \dots$
- assume $\tilde{\theta}$ consistent; regularity; then $\tilde{\theta} \xrightarrow{d} N(\theta, V)$
- $V(\theta) = J^{-1}(\theta)I(\theta)J^{-1}(\theta)$ (sandwich variance)
- $J(\theta) = E_{\theta}\{-\partial U_2(\theta)/\partial \theta\}$
- $I(\theta) = E_{\theta}\{U_2(\theta)U_2(\theta)^T\}$
- $\tilde{\theta}$ not fully efficient, because $J \neq I$, even though components of U_2 may satisfy the Bartlett identities
- **loss of efficiency seems to be small**

Example: symmetric normal

- $Y_i \sim N(0, R)$, $\text{var}(Y_{ir}) = 1$, $\text{corr}(Y_{ir}, Y_{is}) = \rho$
- compound bivariate normal densities to form ℓ_2

$$\ell_2(\rho; y_1, \dots, y_n) = -\frac{nq(q-1)}{4} \log(1-\rho^2) - \frac{q-1+\rho}{2(1-\rho^2)} SS_w - \frac{(q-1)(1-\rho)}{2(1-\rho^2)} \frac{SS_b}{q}$$

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

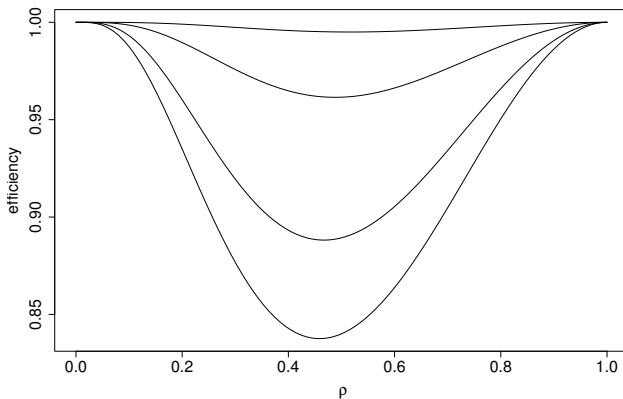
$$\ell(\rho; y_1, \dots, y_n) = -\frac{n(q-1)}{2} \log(1-\rho) - \frac{n}{2} \log\{1+(q-1)\rho\} - \frac{1}{2(1-\rho)} SS_w - \frac{1}{2(\{1+(q-1)\rho\})} \frac{SS_b}{q}$$

... symmetric normal

$$\text{a.var}(\tilde{\rho}) = \frac{2}{nq(q-1)} \frac{(1-\rho)^2}{(1+\rho^2)^2} c(q^2, \rho^4)$$

$$O\left(\frac{1}{n}\right) \quad O(1)$$

$$n \rightarrow \infty \quad q \rightarrow \infty$$



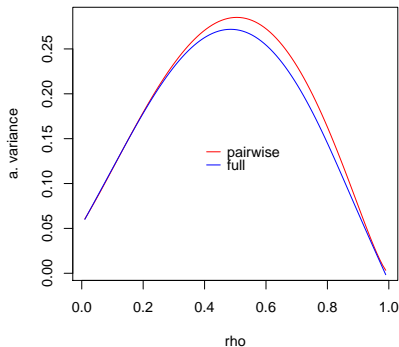
Example: dichotomized MV Normal

$$Y_r = 1\{Z_r > 0\} \quad Z \sim N(0, R)$$

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

$$\text{a.var}(\tilde{\rho}) = \frac{1}{n} \frac{4\pi^2}{q^2} \frac{(1 - \rho^2)}{(q - 1)^2} \text{var}(T) \quad T = \sum_{s < r} (2y_r y_s - y_r - y_s)$$

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



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

Likelihood ratio statistics

- full log-likelihood

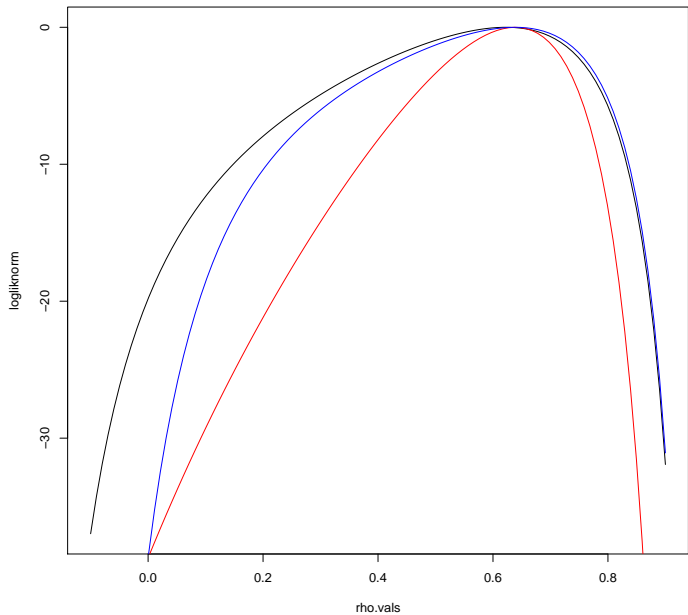
$$w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{d} \chi_p^2$$

- pairwise log-likelihood ($p = 1$)

$$w_{pair}(\theta) \times \frac{J(\theta)}{I(\theta)} \xrightarrow{d} \chi_1^2$$

- $J(\theta) = E_{\theta}\{-\partial U_2(\theta)/\partial\theta\}$
- $I(\theta) = E_{\theta}\{U_2(\theta)U_2(\theta)^T\}$

go to R



n fixed, $q \rightarrow \infty$

- $l_1(\theta; \mathbf{y}) = \sum_s l_1(\theta; y_s)$
- $U_1(\tilde{\theta}) = 0 \simeq U_1(\theta) + (\tilde{\theta} - \theta) \underbrace{U_1'(\theta)}_{q \text{ terms}}$
- $EU_1 = 0, \text{var} U_1 = \sum \text{var} l'_{1r} + \underbrace{\sum \sum \text{cov} l'_{1r} l'_{1s}}_{\text{could be } O(q^2) \text{ terms}}$
- using bivariate: $U_2(\theta) = \sum_{s < r} l'_{rs} - aq \sum_s l'_s$
- $U_2(\tilde{\theta}) = 0 \simeq \sum_{s < r} l'_{st}(\theta) - aq \sum_s l'_s(\theta) + (\tilde{\theta} - \theta) \sum_{s < r} l''_{rs}(\theta) - aq \sum_s l''_s(\theta)$
- $\text{var} : \left\{ \begin{array}{ccc} \text{var} l'_{rs} & - 2aq \text{cov}(l'_{rs} l'_s) & + a^2 q^2 \text{var} l'_s \end{array} \right\}$
- $\begin{array}{ccc} \binom{q}{4} & 2a \binom{q}{3} & a^2 q^2 \binom{q}{2} \end{array}$
- $E: O(q^2)$

... $q \rightarrow \infty$

symmetric normal

$$\text{a.var}(\tilde{\rho}) = \frac{2}{nq(q-1)} \frac{(1-\rho)^2}{(1+\rho^2)^2} c(q^2, \rho^4)$$

$$\begin{array}{ll} O\left(\frac{1}{n}\right) & O(1) \\ n \rightarrow \infty & q \rightarrow \infty \end{array}$$

dichotomized mv normal:

$$\text{a.var}(\tilde{\rho}) = \frac{1}{n} \frac{4\pi^2}{q^2} \frac{(1-\rho^2)}{(q-1)^2} \text{var}(T) \quad T = \sum_{s < r} (2y_r y_s - y_r - y_s)$$

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

Examples where $q \rightarrow \infty$

- single long time series
- spatial models (q indexes spatial sites)
- usually assume decaying correlations, so q can play the role of n
- genetics (McVean et al., 2002; Fearnhead, 2003; Hudson, 2001)
- 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
- pairwise likelihood estimate of recombination rate is not consistent as length of sequence $\rightarrow \infty$ (Fearnhead)
- in this case because score function does not have mean zero
- a more complex version (Fearnhead & Donnelly, 2001) does give consistent estimates

Composite Likelihood

- Lindsay (1988): a general approach $Y \sim f(y; \theta), y \in R^q$
- $L(\theta; y) \propto f(y; \theta)$ $L_{comp}(\theta; y) = \prod_r L_r(\theta; y)$
- each component $L_r(\theta)$ is proportional to a density (conditional or marginal)
- hence has mean zero, etc.
- example with $q = 3$:

$$L(\theta) = f(y_3 | y_2, y_1; \theta) f(y_2 | y_1; \theta) f(y_1; \theta)$$

$$L_{pseudo}(\theta) = f(y_1 | y_2; \theta) f(y_1 | y_3; \theta) f(y_2 | y_3; \theta) f(y_2 | y_1; \theta) \dots$$

$$L_{pairwise}(\theta) = f(y_1, y_2; \theta) f(y_1, y_3; \theta) f(y_2, y_3; \theta)$$

- each component is relatively simple to compute
- resulting estimating equation leads to consistent, but not efficient point estimator

Spatial data

- pseudo-likelihood for spatial data (Besag, 1974)
- auto-normal: $y_s \mid y_{[s]} \sim N(\theta \underline{w}_s^T y, \sigma^2)$
- $W = (\underline{w}_1, \dots, \underline{w}_q)$, $w_{sr} = 1$ if y_s is a neighbour of y_r
- full joint distribution is multivariate normal
- log-likelihood $\ell(\theta) \propto \theta y^T W y / 2 - c(\theta)$
- $c(\theta)$ depends on neighbourhood scheme in a complex way and typically not computable
- Besag suggested using the product $\prod f(y_s \mid y_{[s]})$ as a pseudo-likelihood
- auto-logistic: $y_s \mid y_{[s]} \sim \text{Bernoulli}(p_s)$, logit $p_s = \theta \underline{w}_s^T y$
- similar exponential family form with complex $c(\theta)$

...Spatial data

- $\ell_2(\theta) = \sum \log f(y_s, y_t; \theta) - aq \sum \log f(y_s)$
- $a = 1/2$ gives approximately Besag likelihood
- Nott and Rydén (1999): spatial binary data generated by thresholding an underlying Gaussian random field
- $Y_s = 1\{Z(s) > u\}$ $\{Z(s); s \in R^2\}$ stationary Gaussian random field, mean zero, variance 1, correlation function $R(\cdot; \theta)$
- similar approach in Heagerty and Lele (1998)
- both papers use pairwise likelihood
- dichotomized m.v. normal

Clustered (multi-level) data

- a simplified random effects model
- $Y_{is} = \mu + \xi_i + \varepsilon_{is}$
- $\xi_i \sim N(0, \sigma_\xi^2), \varepsilon_{is} \sim N(0, \sigma_\varepsilon^2)$
- $\rho = \sigma_\xi^2 / (\sigma_\xi^2 + \sigma_\varepsilon^2) \quad \mu = 0, \quad \sigma^2 = 1$
- nested random effects for binary data **Renard et al. (2004)**
- $Pr(y_{is} = 1) = \Phi(\beta_0 + \beta_1 x_{is} + b_{0i} + b_{1i} x_{is})$
- $b_0 \sim N(0, \sigma_0^2), \quad b_1 \sim N(0, \sigma_1^2)$
- comparison of bias and variance via simulation
- reported efficiency loss between 5 and 18 % compared to full m.l. estimation
- estimating variance parameters more difficult than estimating mean parameters

... Example

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

659

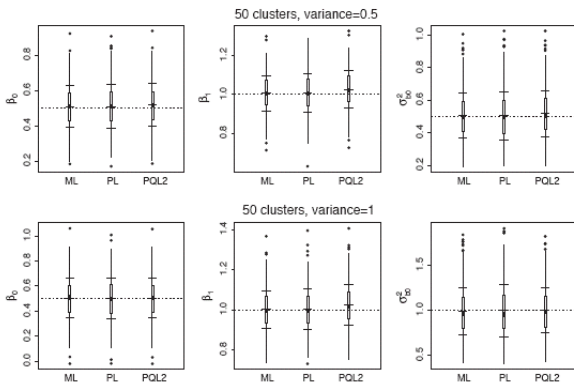


Fig. 2. Boxplots of ML, PL and PQL2 simulated parameter estimates under Model (9) with random intercept $\sim N(0, \sigma_{b_0}^2)$. Top panel: 50 clusters with $\sigma_{b_0}^2 = 0.5$; Bottom panel: 50 clusters with $\sigma_{b_0}^2 = 1$.

... mixed models

- Bellio and Varin (2005): generalized linear mixed models; crossed random effects; emphasis on bias
- $g\{E(Y_{is} | u_i, v_s)\} = x_{is}^T \beta + u_i + v_s$
- $u_i \sim N(0, \sigma_u^2), \quad v_s \sim N(0, \sigma_v^2)$
- full likelihood $L(\theta; y)$ requires integration over $R^{q_1} \times R^{q_2}$
- pairwise likelihood $L_{pair}(\theta; y) = \prod_{i=1}^n \prod_{s < r} f(y_{ir}, y_{is})$ uses only integrals in R^3
- Varin, Host and Skare (2004): generalized linear mixed models

... mixed models

- Zhou and Joe (2005): familial data
- binary data for clusters (Kuk and Nott, 2000)
- $Y_i = (Y_{i1}, \dots, Y_{iq_i})$ observations within a cluster; n clusters
- each Y_{is} Bernoulli with, e.g. logit $Pr(Y_{is} = 1) = x_{is}^T \beta$
- observations within a cluster assumed to have a pairwise dependence of the form
- $\log \frac{P_{i,sr}(1, 1)P_{i,sr}(0, 0)}{P_{i,sr}(1, 0)P_{i,sr}(0, 1)} = z_{isr}^T \alpha$, etc.
- or could model the correlation coefficient directly
- higher order dependencies not explicitly modelled

Some questions

- $U_2(\theta) = \sum_{i=1}^n \sum_{s < r} \ell'_{rs}(\theta) - aq \sum \ell'_s(\theta)$
- $\text{var} \sum_{i=1}^n \sum_{r < s} \ell'_{rs}(\theta) \sim nq^4$
- sum is thus $O_p(q^2)$, of same order as information term
- as $q \rightarrow \infty$ can the q^4 term be eliminated by choice of a ?
- as $n \rightarrow \infty$ can a be chosen to maximize efficiency?
(Lindsay, 1988)
- $a_{opt} = \frac{E(\ell'_r \ell'_s) E(-\ell''_{rs}) + E(\ell'_{rs})^2 E(-\ell''_s)}{E\ell'_s{}^2 E(-\ell''_{rs}) + E(\ell'_s \ell''_{rs}) E(-\ell''_s)}$
- $\text{var}U$ is minimized at $a = E(\ell'_s \ell'_{rs}) / E(\ell'_s)^2$ (depends on θ)
- different weighting adjustments used in Kuk and Nott, Parner, Renard et al (weighting by cluster sizes)

Relation to GEE

- same if $Y_i \sim N(\mu_i, V_i)$, $V_i = \text{diag}(\sigma_{is}^2)$
- $g(\mu_{is}) = x_{is}^T \beta$, $\sigma_{ir}^2 = \phi h(\mu_{is})$
- pairwise score proportional to GEE under independence
- GEE fully efficient if correlations nonzero
- multivariate binary data: lead to same score equations under independence
- PL is fully efficient if $\rho_{ir} \neq 0$, $\rho_{irs} \dots$ all zero

Some related work

- Liang and Yu (2003): network tomography
- (X_1, \dots, X_J) hidden variables measuring network performance; X_i i.i.d. $\sim f(\cdot; \theta)$
- (Y_1, \dots, Y_I) observed measurements ($I \ll J$)
- model $Y = AX$, where A is routing matrix of 1's and 0's (not full rank)
- pairwise likelihood constructed by analysing all possible pairs of rows in A
- efficient, easier to maximize

... Some related work

- Vidoni and Varin (2005): state space models $\{X_n, Y_{n+1}\}_{n \geq 0}$
- observation equation $Y_n = z_n(X_n, V_n)$
- transition equation $X_n = u_n(X_{n-1}, W_n)$ (HMM)
- Hjort and Vidoni (2005): Markov chains on finite state spaces
- Song, Fan and Kalbfleisch (2005): maximization by parts
- Hu and Zidek (2002): relevance weighted likelihood

Questions

- is PL useful for modelling when no joint distribution is available (and may not exist?)
- e.g. extreme values, survival data (Parner, 2001)
- can any progress be made in choosing a or in other weighting schemes for $q \rightarrow \infty$
- asymptotic theory in n, q 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

...Questions

- how to investigate robustness systematically
- other generalizations of likelihood ideas important for applications
- e.g. semiparametric likelihood, empirical likelihood, MCMC estimated likelihoods