

# Likelihood-based inference in complex models

Nancy Reid

University of Florida

January 13, 2006

David Cox, Grace Yun Yi (U Waterloo)

## Likelihood inference

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

$$\begin{aligned}(\hat{\theta} - \theta)^T \{-\ell''(\hat{\theta})\}(\hat{\theta} - \theta) &\sim \chi_p^2 \\ \ell'(\theta)^T \{-\ell''(\hat{\theta})\}^{-1} \ell'(\theta) &\sim \chi_p^2 \\ 2\{\ell(\hat{\theta}) - \ell(\theta)\} &\sim \chi_p^2\end{aligned}$$

- ▶ asymptotic theory requires  $\hat{\theta} \xrightarrow{P} \theta$
- ▶  $\ell'(\theta)$  behaves like a sum of independent components
- ▶  $E(\ell' \ell'^T) = E(-\ell'')$  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: e.g. mean/variance specifications as in quasi-likelihood; or generalized estimating equations (GEE) with ‘working’ covariance structure

## Examples

- ▶ pseudo-likelihood for spatial data (Besag, 1974)
- ▶ auto-normal:  $y_s \mid y_{[s]} \sim N(\beta \underline{w}_s^T y, \sigma^2)$
- ▶  $W = (\underline{w}_1, \dots, \underline{w}_q)$ ,  $w_{sr} = 1$  if  $y_s$  is a neighbour of  $y_r$
- ▶ full joint distribution is multivariate normal
- ▶ log-likelihood  $\beta y^T W y / 2 - c(\beta)$
- ▶ auto-logistic:  $y_s \mid y_{[s]} \sim \text{Bernoulli}(p_s)$ , logit  $p_s = \beta \underline{w}_s^T y$
- ▶ full joint density is again of exponential family form; again  $c(\beta)$  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
- ▶ Lindsay (1988) studied the properties of a more general form under the name composite likelihood

## ...Examples

- ▶ dichotomized multivariate normal (latent variables)
- ▶  $Z = (Z_1, \dots, Z_q)$  multivariate normal, mean 0, variance 1, correlation matrix  $R$
- ▶  $Y_s = 1\{Z_s > 0\}$
- ▶ joint density of  $Y = (Y_1, \dots, Y_q)$  involves  $q$ -dimensional normal integral
- ▶ hard to compute, results may not be robust to specification for  $Z$
- ▶ spatial binary data generated this way in Nott and Ryden (1999) by thresholding an underlying Gaussian random field

## ...Examples

- ▶ binary data for clusters, or... (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

## Composite likelihood

- ▶ Besag's pseudo-likelihood compounds a set of conditional densities
- ▶ pairwise likelihood (Hjort, unpublished; Nott and Ryden, 1999): compound a set of marginal densities for dependent observations taken in pairs
- ▶ composite log likelihood (Lindsay): compound components which are individually likelihoods, i.e. proportional to densities for observations, conditional, marginal or ...
- ▶ example with  $q = 3$ :

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

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

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

## 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 $\theta$ 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  $\ell_2$

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

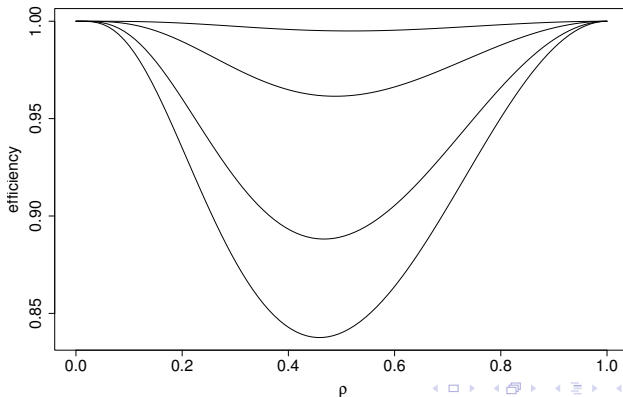
$$\text{SS}_w = \sum_{i=1}^n \sum_{s=1}^q (y_{is} - \bar{y}_i.)^2, \quad \text{SS}_b = \sum_{i=1}^n y_i^2$$

$$\ell(\rho; y_1, \dots, y_n) = -\frac{n(q-1)}{2} \log(1 - \rho) - \frac{n}{2} \log\{1 + (q-1)\rho\} - \frac{1}{2(1-\rho)} \text{SS}_w - \frac{1}{2\{1 + (q-1)\rho\}} \frac{\text{SS}_b}{q}$$

(2)

## ...symmetric normal

$$\text{a.var}(\tilde{\rho}) = \frac{2}{nq(q-1)} \frac{(1-\rho)^2}{(1+\rho^2)^2} c(q^2, \rho^4)$$
$$O\left(\frac{1}{n}\right) \quad O(1)$$
$$n \rightarrow \infty \quad q \rightarrow \infty$$



## ...symmetric normal

- ▶ special case of simplified random effects model
- ▶  $Y_{is} = \mu + \xi_i + \varepsilon_{is}$
- ▶  $\xi_i \sim N(0, \sigma_\xi^2), \varepsilon_{is} \sim N(0, \sigma_\varepsilon^2)$
- ▶  $\rho = \sigma_\xi^2 / (\sigma_\xi^2 + \sigma_\varepsilon^2)$
- ▶ pairwise likelihood for more complex models with random effects in
  1. Renard et al. (2004): clustered binary data (probit link)
  2. Bellio and Varin (2005): crossed random effects
  3. Varin, Host and Skare (2004): generalized linear mixed models
  4. Henderson and Shimakura (2003): longitudinal count data (Poisson-gamma frailty)
  5. Zhou and Joe (2005): familial data
  6. Parner (2001): bivariate survival data (frailty model)

## Example: Renard et al. 2004

- ▶ random effects model for binary data using probit link
- ▶  $Pr(y_{is} = 1) = \Phi(\beta_0 + \beta_1 x_{is} + b_{0i} + b_{1i} x_{is})$
- ▶  $b_0 \sim N(0, \sigma_0^2)$ ,  $b_1 \sim N(0, \sigma_1^2)$
- ▶ reported efficiency loss between 5 and 18 % compared to full m.l. estimation
- ▶ estimating variance parameters more difficult than estimating mean parameters
- ▶ univariate marginals not used; could efficiency improve using these?



## Example: dichotomized MV Normal

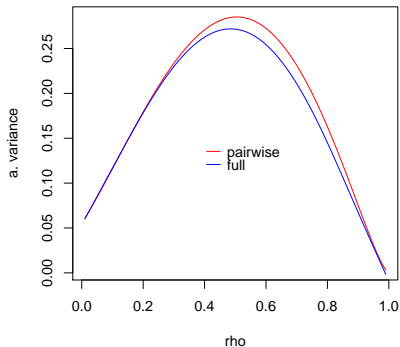
$$Y_r = 1\{Z_r > 0\}, Z \sim N(0, R)$$

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

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

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

## ...dichotomized mvn



$\rho$	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
$\rho$	0.60	0.70	0.80	0.90	0.95	0.98

## Further consideration of $q \rightarrow \infty$

- ▶  $l_1(\theta; \mathbf{y}) = \sum_s l(\theta; y_s)$
- ▶  $l'_1(\tilde{\theta}) = 0 \simeq l'_1(\theta) + (\tilde{\theta} - \theta) \underbrace{l''_1(\theta)}_{q \text{ terms}}$
- ▶  $E l'_1 = 0, \text{var} l'_1 = \sum \text{var} l'_{1r} + \underbrace{\sum \sum \text{cov} l'_{1r} l'_{1s}}_{\text{could be } O(q^2) \text{ terms}}$
- ▶ using bivariate:  $U_2(\theta) = \sum_{s < r} l'_{rs} - aq \sum_s l'_s$
- ▶  $U_2(\tilde{\theta}) = 0 \simeq$   
 $\sum_{s < r} l'_{st}(\theta) - aq \sum_s l'_s(\theta) + (\tilde{\theta} - \theta) \sum_{s < r} l''_{rs}(\theta) - aq \sum_s l''_s(\theta)$
- ▶ **var** :  $\{\text{var} l'_{rs} - 2aq \text{cov}(l'_{rs} l'_s) + a^2 q^2 \text{var} l'_s\}$
- ▶  $\binom{q}{4} \quad 2a \binom{q}{3} \quad a^2 q^2 \binom{q}{2}$
- ▶ **E**:  $O(q^2)$

## 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)
- ▶ as  $n \rightarrow \infty$  can show a.  $\text{var}(\tilde{\theta})$  minimized by choosing  $a$  as a function of variances and covariances:
- ▶ 
$$a_{opt} = \frac{E(\ell'_r \ell'_s) E(-\ell''_{rs}) + E(\ell'_{rs})^2 E(-\ell''_s)}{E\ell'_s{}^2 E(-\ell''_{rs}) + E(\ell'_s \ell''_{rs}) E(-\ell''_s)}$$
- ▶  $\text{var}U$  is minimized at  $a = E(\ell'_s \ell'_{rs}) / E(\ell'_s)^2$  (depends on  $\theta$ )
- ▶ 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

## Conclusions

- ▶ is PL useful for modelling when no joint distribution is available (and may not exist?)
- ▶ e.g. extreme values, survival data
- ▶ 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, as in GEE1, GEE2

## ...Conclusions

- ▶ large  $q$  asymptotics used in estimation of recombination rate from DNA sequences (Fearnhead, 2003; McVean et al 2002; Hudson 2001) (also Geyer and Thompson, 1992, 1995)
- ▶ asymptotic and finite sample behaviour of pairwise likelihood ratio tests; large  $q$  and large  $n$
- ▶ how to investigate robustness systematically
- ▶ other generalizations of likelihood ideas important for applications
- ▶ e.g. semiparametric likelihood (Murphy and Van der Waart), empirical likelihood, etc.

## References

...coming

but start with Christiano Varin's home page at

`www.stat.unipd.it`