Likelihood-based inference in complex models

Nancy Reid

University of Toronto

January 26, 2006

David Cox, Grace Yun Yi (U Waterloo)
Likelihood inference

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

\[
(\hat{\theta} - \theta)^T \{-\ell''(\hat{\theta})\} (\hat{\theta} - \theta) \sim \chi_p^2
\]

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

\[
2\{\ell(\hat{\theta}) - \ell(\theta)\} \sim \chi_p^2
\]

- asymptotic theory requires \( \hat{\theta} \xrightarrow{p} \theta \quad \ell'(\hat{\theta}) = 0, \quad E\ell'(\theta) = 0 \)
- a central limit theorem for \( \ell'(\theta) \)
- \( E(\ell'\ell'^T) = E(-\ell'\ell'') \) 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
Composite Likelihood

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

$$L(\theta) = f(y_3 \mid y_2, y_1; \theta)f(y_2 \mid y_1; \theta)f(y_1; \theta)$$
$$L_{\text{pseudo}}(\theta) = f(y_1 \mid y_2; \theta)f(y_1 \mid y_3; \theta)f(y_2 \mid y_3; \theta)f(y_2 \mid y_1; \theta)\ldots$$
$$L_{\text{pairwise}}(\theta) = f(y_1, y_2; \theta)f(y_1, y_3; \theta)f(y_2, y_3; \theta)$$

- 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: $y_s \mid y_s \sim N(\theta w_s^T y, \sigma^2)$
- $W = (w_1, \ldots, w_q)$, $w_{sr} = 1$ if $y_s$ is a neighbour of $y_r$
- full joint distribution is multivariate normal
- log-likelihood $\ell(\theta) \propto \theta y^T Wy / 2 - c(\theta)$
- $c(\theta)$ 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: $y_s \mid y_s \sim Bernoulli(p_s)$, logit $p_s = \theta w_s^T y$
- similar exponential family form with complex $c(\theta)$

plausible models exist but difficult to evaluate
...Spatial data

- Nott and Rydén (1999): spatial binary data generated by thresholding an underlying Gaussian random field
- \( Y_s = 1\{Z(s) > u\} \quad \{Z(s); s \in R^2\} \) stationary Gaussian random field, mean zero, variance 1, correlation function \( R(\cdot; \theta) \)
- pairwise likelihood \( L_{\text{pair}}(\theta; y) = \prod_{s,r} f(y_s, y_r) \)
- Heagerty and Lele (1998): thresholding random field, regression structure in the mean
- pairwise likelihood with slightly different parameterization of \( f(y_s, y_r) \)
- score equations \( U(\theta; y) = \sum_{s,r} w_{s,r} \log f(y_s, y_r) \)
- model leads to score components involving \( y_s - EY_s \) and \( (y_s - EY_s)(y_r - Ey_r) - \sigma_{r,s} \)

plausible models exist but may not be reliable
Models for multilevel data

- binary data for clusters (Kuk and Nott, 2000)
- \( Y_i = (Y_{i1}, \ldots, Y_{iq_i}) \) 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
  \[ \log \frac{P_{i,sr}(1,1)P_{i,sr}(0,0)}{P_{i,sr}(1,0)P_{i,sr}(0,1)} = z_{isr}^T \alpha, \text{ etc.} \]
- or could model the correlation coefficient directly
- higher order dependencies not explicitly modelled

simpler models preferred
Generalized linear mixed models

- crossed random effects: \( g\{E(Y_{is} \mid u_i, v_s)\} = x_{is}^T \beta + u_i + v_s \)
  Bellio & Varin: 2005
- \( u_i \sim N(0, \sigma_u^2), \quad v_s \sim N(0, \sigma_v^2) \)
- 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 \( R^3 \)
- nested random effects for binary data Renard et al. (2004)
- \( Pr(y_{is} = 1) = \Phi(\beta_0 + \beta_1 x_{is} + b_{0i} + b_{1i} x_{is}) \)
- \( b_0 \sim N(0, \sigma_0^2), \quad b_1 \sim N(0, \sigma_1^2) \)
- frailty model for counts \( Y_{is} \sim Poisson(Z_{is} \exp(\alpha_{is} + \beta x_i)) \)
  Henderson & Shimakura (2003)
- \( Z_i = (Z_{i1}, \ldots, Z_{iq}) \) follows multivariate Gamma; to be integrated out to form full likelihood
- use instead pairwise likelihood as more tractable
Pseudo-likelihood

(Cox and R, 2004)

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

\[
\ell_2(\theta; y) = \sum_{r<s} \log f_{rs}(y_r, y_s) - aq \sum_s \log f_s(y_s) \\
= \sum_{r<s} \log f_{rs}(y_r, y_s) - aq \ell_1(\theta; y)
\]

- \( n \) independent observations:

\[
\ell_2(\theta; y_1, \ldots, y_n) = \sum_{i=1}^n \ell_2(\theta; y_i)
\]

- score function \( U_2(\theta; y_1, \ldots, y_n) = \frac{\partial \ell_2(\theta; y_1, \ldots, y_n)}{\partial \theta} \)
Estimation of $\theta$ from $U_2$

- $U_2(\tilde{\theta}) = 0$

- $U_2(\tilde{\theta}) = U_2(\theta) + (\tilde{\theta} - \theta)^T \frac{\partial U_2(\tilde{\theta})}{\partial \theta} + \ldots$

- 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 \( \ell_2 \)

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

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

\[
\ell(\rho; y_1, \ldots, y_n) = -\frac{n(q-1)}{2} \log(1 - \rho) - \frac{n}{2} \log\{1 + (q - 1)\rho\}
- \frac{1}{2(1 - \rho)} SS_w - \frac{1}{2\{1 + (q - 1)\rho\}} \frac{SS_b}{q}
\]
...symmetric normal

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

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

\[ n \rightarrow \infty \quad q \rightarrow \infty \]
...symmetric normal

- special case of simplified random effects model
- \( Y_{is} = \mu + \xi_i + \varepsilon_{is} \)
- \( \xi_i \sim N(0, \sigma_\xi^2), \varepsilon_{is} \sim N(0, \sigma_\varepsilon^2) \)
- \( \rho = \sigma_\xi^2 / (\sigma_\xi^2 + \sigma_\varepsilon^2) \)
- comparison of bias and variance via simulation in
- Renard et al. (2004): clustered binary data (probit link)
- reported efficiency loss between 5 and 18 % compared to full m.l. estimation
- estimating variance parameters more difficult than estimating mean parameters
- Bellio and Varin (2005): crossed random effects; emphasis on bias
- Varin, Host and Skare (2004): generalized linear mixed models
- Zhou and Joe (2005): familial data
Fig. 2. Boxplots of ML, PL and PQL2 simulated parameter estimates under Model (9) with random intercept $\sim N(0, \sigma^2_{\beta_0})$. Top panel: 50 clusters with $\sigma^2_{\beta_0} = 0.5$; Bottom panel: 50 clusters with $\sigma^2_{\beta_0} = 1$. 
Example: dichotomized MV Normal

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

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

\[ + (1 - y_r) y_s \log P_{01} + (1 - y_r)(1 - y_s) \log P_{00} \} \]

\[ \text{a.var}(\tilde{\rho}) = \frac{1}{n} \frac{4\pi^2}{q^2} \frac{(1 - \rho^2)}{(q - 1)^2} \text{var}(T) \]

\[ T = \sum_{s<r} (2y_r y_s - y_r - y_s) \]

\[ \text{var}(T) = q^4(p_{1111} - 2p_{111} + 2p_{11} - p_{11}^2 + \frac{1}{4}) + q^3(-6p_{1111}...) + q^2(...) + q(...) \]
Further consideration of $q \to \infty$

- $\ell_1(\theta; y) = \sum_s \ell(\theta; y_s)$
- $\ell'_1(\tilde{\theta}) = 0 \approx \ell'_1(\theta) + (\tilde{\theta} - \theta) \ell''_1(\theta)$
  \[ q \text{ terms} \]
  \[
  E\ell'_1 = 0, \quad \text{var}\ell_1' = \sum \text{var}\ell'_1r + \sum \sum \text{cov}\ell'_1r\ell'_1s
  \]
  \[ \text{could be } O(q^2) \text{ terms} \]
- using bivariate: $U_2(\theta) = \sum_{s<r} \ell'_rs - aq \sum_s \ell'_s$
- $U_2(\tilde{\theta}) = 0 \approx \sum_{s<r} \ell'_st(\theta) - aq \sum_s \ell'_s(\theta) + (\tilde{\theta} - \theta) \sum_{s<r} \ell''_rs(\theta) - aq \sum_s \ell''_s(\theta)$

- $\text{var}: \{\text{var}\ell'_rs - 2aq\text{cov}(\ell'_rs \ell'_s) + a^2 q^2 \text{var}\ell'_s\}$
  \[
  \left(\frac{q}{4}\right) \quad 2a\left(\frac{q}{3}\right) \quad a^2 q^2 \left(\frac{q}{2}\right)
  \]
- $E: O(q^2)$
... \( q \rightarrow \infty \)

symmetric normal

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

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

\[
n \rightarrow \infty \quad q \rightarrow \infty
\]

dichotomized mv normal:

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

\[
\text{var}(T) = q^4 (p_{1111} - 2p_{111} + 2p_{11} - p_{11}^2 + \frac{1}{4}) + q^3 (-6p_{1111}) + q^2 (...) + q (...)\]
... $q \rightarrow \infty$

- spatial models ($q$ indexes spatial sites)
- usually assume decaying correlations, so $q$ can play the role of $n$
- similarly with single long time series
- 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
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:
- $a_{opt} = \frac{E(\ell'_s \ell'_r)E(-\ell''_{rs}) + E(\ell'_{rs})^2 E(-\ell''_s)}{E\ell'_s E(-\ell''_{rs}) + E(\ell'_s \ell''_{rs}) E(-\ell''_s)}$
- $\text{var} U$ is minimized at $a = E(\ell'_s \ell'_r)/E(\ell'_s)^2$ (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 $\rho_{ir} \neq 0, \rho_{irs} \ldots \text{all zero}$
- log-likelihood ratio statistic based on pairwise likelihood has limiting $\chi^2$ distribution ($n \to \infty$)
Some related work

- Liang and Yu (2003): network tomography
- \((X_1, \ldots, X_J)\) hidden variables measuring network performance; \(X_i \text{ i.i.d. } \sim f(\cdot; \theta)\)
- \((Y_1, \ldots, Y_I)\) observed measurements \((I << J)\)
- 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)
- Song, Fan and Kalbfleisch (2005): maximization by parts
- Neal, Li and Zhang: feature selection. Likelihood constructed from part of the data; information that remaining data was excluded is incorporated
- 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 $q \to \infty$
• 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