Likelihood inference in complex models

Nancy Reid

December 6, 2009

with Cristiano Varin and David Firth
Ca' Foscari University, Venice and University of Warwick

2009 CMS WINTER MEETING
December 5 - 7, Windsor, Ontario

The setup

- Data: $y = (y_1, \ldots, y_n)$ $\quad x_1, \ldots, x_n \quad i = 1, \ldots, n$
- Model for the probability distribution of y_i given x_i
- Density (with respect to, e.g., Lebesgue measure)
- $f(y_i \mid x_i) \qquad f(y \mid x) > 0, \int f(y \mid x) dy = 1$
- joint density for $y = f(y \mid x) = \prod f(y_i \mid x_i)$ independence

- parameters for the density $f(y \mid x; \theta)$, $\qquad \theta = (\theta_1, \ldots, \theta_d)$
- often $\theta = (\psi, \lambda)$
- θ could have dimension $d > n$ (e.g. genetics)
- θ could have infinite dimension e.g. $E(y \mid x) = \theta(x)$ 'smooth'

Definitions

- Likelihood function

$$L(\theta; y) = L(\theta; y_1, \ldots, y_n) = f(y_1, \ldots, y_n; \theta) = \Pi_{i=1}^{n} f(y_i; \theta)$$

- Log-likelihood function:

$$\ell(\theta; y) = \log L(\theta; y)$$

- Maximum likelihood estimator (MLE)

$$\hat{\theta} \quad = \quad \arg\sup_{\theta} L(\theta; y) \qquad \hat{\theta}(y)$$

Example: time series studies of air pollution[1]

- y_i: number of deaths in Toronto due to cardio-vascular or respiratory disease on day i
- x_i: 24 hour average of PM_{10} or O_3 in Toronto on day i, maximum temperature, minimum temperature, dew point, relative humidity, day of the week, ...
- model: Poisson distribution for counts
- $$f(y_i;\theta) = \{\mu_i(\theta)\}^{y_i} \exp\{-\mu_i(\theta)\}$$
- $$\log \mu = \alpha + \psi PM_{10} + S(time, df_1) + S(temp, df_2)$$
- $S(time, df_1)$ a 'smooth' function
- typically $S(\cdot, df_1) = \sum_{j=1}^{df_1} \lambda_j B_j(\cdot)$
- $B_j(\cdot)$ known basis functions   usually splines
- $\theta = (\alpha, \psi, \lambda_1, \lambda_2)$ with dimension $df_1 + df_2 + 2$

[1] Peng et al., 2006

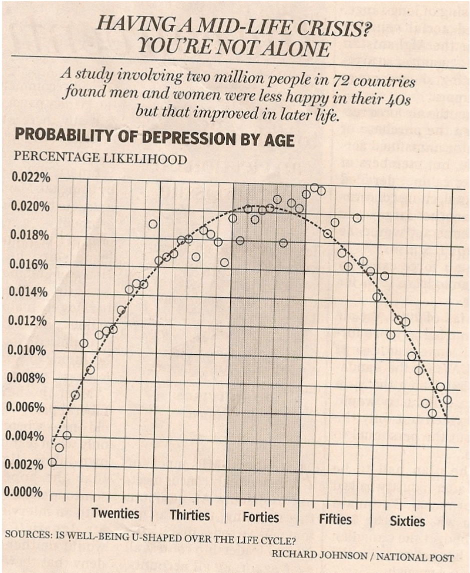
Example: longitudinal study of migraine sufferers[2]

- latent variable $Y_{ij}^* = x_{ij}^T \beta + U_i + \epsilon_{ij}$
- observed variable $y_{ij} \in \{1, \dots, h\} \leftrightarrow \alpha_{y_{ij}-1} < Y_{ij}^* < \alpha_{y_{ij}}$,
- e.g. no headache, mild, moderate, ... intense
- x_{ij} covariates – age, education, change in barometric pressure, use of painkillers, ...
- U_i, ϵ_{ij} random effects between and within subjects
- $\epsilon_{ij} = \rho \epsilon_{i,j-1} + (1-\rho^2)^{1/2} \eta_{ij}$, serial correlation over time
- $$L(\theta; y) = \prod_{i=1}^{n} \int \cdots \int \phi_m(z_{i1}, \dots, z_{im_i}; R) dz_{i1} \dots dz_{im_i}$$
- $R_{ij} = (\sigma^2 + \rho^{|i-j|})/(\sigma^2 + 1)$

[2] Czado & Varin, 2010

Very widely used

- IEEE Transactions on Computational Biology and Bioinformatics
- Crop Breeding, Genetics and Cytology
- The Review of Financial Studies
- IEEE Transactions on Information Theory
- Journal of the American Medical Association
- Molecular Biology and Evolution
- Physical Review D
- US Patent Office



National Post, Toronto, Jan 30 2008

Composite likelihood

- Model: $Y \sim f(y;\theta), \quad Y \in \mathcal{Y} \subset \mathbb{R}^p, \quad \theta \in \mathbb{R}^d$
- 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
- single p-dimensional response $Y = Y_i$

Composite conditional likelihood

- Pseudo-likelihood — Besag, 1974

$$CL(\theta; y) = \prod_{r=1}^{p} f(y_r \mid \{y_s : y_s \text{ a neighbour of } y_r\}; \theta)$$

- or use blocks of observations — Vecchia, 1988; Stein et al., 2004
- stratified case-control studies — Liang, 1987

$$CL(\theta; y) = \prod_{r=1}^{p} \prod_{s=r+1}^{p} f(y_r \mid y_r + y_s; \theta)$$

- pairwise conditional $CL(\theta; y) = \prod_{r=1}^{p} \prod_{s=1}^{p} f(y_r \mid y_s; \theta)$
- full conditional $CL(\theta; y) = \prod_{r=1}^{p} f(y_r \mid y_{(r)}; \theta)$

Molenberghs & Verbeke, 2005

Composite marginal likelihood

- $$CL(\theta; y) = \prod_{s \in \mathcal{S}} f_s(y_s; \theta), \quad \text{subvectors}$$
- 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: $\prod_{r=1}^{p-1} \prod_{s=r+1}^{p} f(y_r - y_s; \theta)$

  Curriero & Lele, 1999
- and even mixtures of *CCL* and *CML*

Derived quantities

- log composite likelihood: $c\ell(\theta; y) = \log CL(\theta; y)$
- score function: $U(\theta; y) = \nabla_\theta c\ell(\theta; y) = \sum_{s \in \mathcal{S}} U_s(\theta; y)$
  $E\{U(\theta; Y)\} = 0$
- maximum composite likelihood est.: $\hat{\theta}_{CL} = \arg\sup c\ell(\theta; y)$
  $U(\hat{\theta}_{CL}) = 0$

- variability matrix: $J(\theta) = var_\theta\{U(\theta; Y)\}$
- 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: $Y_1, \ldots, Y_n$ $\qquad CL(\theta; y) = \prod_{i=1}^{n} CL(\theta; y_i)$
- $$\sqrt{n}(\hat{\theta}_{CL} - \theta) \dot{\sim} N\{0, G^{-1}(\theta)\} \qquad G(\theta) = H(\theta)J(\theta)^{-1}H(\theta)$$
- $w(\theta) = 2\{c\ell(\hat{\theta}_{CL}) - c\ell(\theta)\} \dot{\sim} \sum_{a=1}^{d} \mu_a Z_a^2 \quad Z_a \sim N(0,1)$
- $\mu_1, \ldots, \mu_d$ eigenvalues of $J(\theta)H(\theta)^{-1}$
- $w(\psi) = 2\{c\ell(\hat{\theta}_{CL}) - c\ell(\tilde{\theta}_\psi)\} \dot{\sim} \sum_{a=1}^{d_0} \mu_a Z_a^2$
- constrained estimator: $\tilde{\theta}_\psi = \arg\sup_{\theta=\theta(\psi)} c\ell(\theta; y)$
- $\mu_1, \ldots, \mu_{d_0}$ eigenvalues of $(H^{\psi\psi})^{-1} G^{\psi\psi}$
-

Kent, 1982

Many 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
- computer experiments with high-dimensional Gaussian process ($n = 20,000$) — Bingham, 2009
- spatial extremes — Padoan et al. 2009

Point estimation

- $\hat{\theta}_{CL} \dot{\sim} N\{\theta, G^{-1}(\theta)\}$
- $G(\theta) = H(\theta)J(\theta)^{-1}H(\theta)$
- how does this compare to the competition?
- $\hat{\theta}_{ML} \dot{\sim} N\{\theta, I(\theta)^{-1}\}, \quad I(\theta)$ Fisher info matrix
- compare $I(\theta)$ to $G(\theta)$
- analytical calculation or simulation estimates
- compare empirical variances in simulations
- investigate choice of weights for improved efficiency

  Lindsay, 1988; Joe & Lee, 2009
- most natural in context of clustered or longitudinal data $y_i = (y_{i1}, \ldots, y_{im_i})$

Some results on efficiency

- in clusters, use weights

$$\frac{1}{(n_i - 1)\{1 + 0.5(n_i - 1)\}}$$

Joe & Lee, 2009

- or treat parameters in the mean differently from association parameters
- for example using optimal score functions for the parameters in the mean, and CL for association parameters Kuk, 2007
- in time series applications, downweight observations that are far apart in time Joe & Lee, 2009; Varin & Vidoni, 2006

Inference functions

- potential advantage over defining estimating equations directly (GEE)
- $w(\psi) = 2\{c\ell(\hat{\theta}_{CL}) - c\ell(\tilde{\theta}_\psi)\} \dot{\sim} \sum_{a=1}^{d_0} \mu_a Z_a^2$
- approximation by matching first moment or first two moments
- or by saddlepoint approximation
- or by rescaling $w(\psi)$ Chandler & Bate, 2007; Pace et al., 2009
- use in model selection and model averaging

Model selection

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

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

- Bayesian information criterion — Gao and Song, 2009

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

- effective number of parameters

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

- model averaging — Hjort and Claeskens, 2008
- selection of tuning parameters in Lasso — Gao and Song, 2009

Some special cases

- Example: multivariate normal:
- $Y \sim N(\underline{\mu}, \Sigma)$: pairwise likelihood estimates $\equiv$ mles
- $Y \sim N(\mu\underline{1}, \sigma^2 R), R_{ij} = \rho$: pairwise likelihood est. $\equiv$ mles
- $Y \sim N(\mu\underline{1}, R)$: loss of efficiency (although small for $\rho > 0$)

- closed exponential families
- $f(y;\theta) = \exp\{\theta^T t(y) - c(\theta)\} = f(t_{A;B} \mid t_B;\theta) f(t_B;\theta)$
- require θ to separate in conditional and marginal pieces
- leads to $\hat{\theta}_{CL} = \hat{\theta}$ and full efficiency
- multivariate vonMises distribution
- Mardia et al., 2008, 2009

Markov chains [3]

- comparison of likelihood

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

- adjoining pairs CML

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

- composite conditional likelihood (= Besag's PL)

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

[3] Hjort and Varin, 2008

... Markov chain example

- Random walk with p 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 composite conditional likelihood (solid line)

Time series and state space models

- $f(y_t, \ldots, y_1; \theta) = f(y_t \mid y_{(t-1)}; \theta) f(y_{t-1} \mid y_{(t-2)}; \theta) \ldots f(y_1; \theta)$
- proposal by Azzalini, 1983: replace $f(y_{t-j} \mid y_{(t-j+1)}; \theta)$ by $f(y_{t-j} \mid y_{t-j+1}; \theta)$
- a version of composite conditional likelihood
- pairwise likelihood $\prod_{s<t} f(y_t, y_s; \theta)$
- more natural to down-weight, or ignore, pairs with $|t - s| > m$
- simplest example, $AR(1)$ with $m = 1$; pairwise likelihood asymptotically fully efficient — Jin, 2009
- efficiency decreases with increasing m — Davis & Yau, 2009
- extension to $AR(1)$ with additive noise (and more)
-

$$
\begin{aligned}
y_t &= \mu + x_t + \epsilon_t \\
x_t &= \gamma x_{t-1} + \eta_t
\end{aligned}
$$

Varin & Vidoni, 2009

Spatial data

- composite conditional likelihood more natural
- but composite marginal likelihood can have better performance
- if the margins are carefully chosen

- Lele & Taper, 2002: $\prod_{i<j} f(y_i - y_j; \theta)$
- reproduces REML for Gaussian case
- better than maximum likelihood

Aspects of robustness

- model robustness
- univariate and bivariate margins only, for example
- means, variances, association parameters
- similar in flavour to generalized estimating equations
- specify lower order distributions, instead of lower order moments
- if there are several joint distributions with the same lower dimensional margins, inference will be robust over that class
- but are there?

... aspects of robustness

- simulation under the wrong model
- example: binary data with higher order correlations simulated
- model with only mean and pairwise correlations fitted
- pairwise likelihood continues to have good efficiency

  Jin, 2009

- example: sparse clustered binary data
- fitted model has wrong correlation structure
- composite conditional likelihood continues to have high efficiency Wang & Williamson, 2005

... aspects of robustness

- 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
- example: hidden Markov model for transitions between N genes
- in principle requires estimation of $2^N \times 2^N$ matrix
- pairwise likelihood reduces computation to $O(N^2)$

Song and Gao, 2009

Missing data

- binary responses — Yi, Zeng and Cook, 2009
- $(y_{ij}, y_{ik}, r_{ij}, r_{ik})$: r_{ij} records missing (0) or not (1)
- generalization to more flexible mean functions (non-parametric) — He & Yi, 2009

Questions about inference

- When Is composite marginal likelihood preferred to conditional composite likelihood ? (always?)
- why is composite likelihood seemingly so efficient?
- where are the exceptions?
- model classes that lead to asymptotic efficiency?

  Mardia et al, 2009
- role of sufficiency?

... questions

- 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
- estimation of eigenvalues of $(H^{\psi\psi})^{-1}G^{\psi\psi}$

... questions

- approximation of distribution of $w(\psi) \dot{\sim} \sum \mu_a Z_a^2$
- Satterthwaite type? $(f\chi^2_d)$: Geys et al, 1999
- saddlepoint approximation?: Kuonen, 2004
- direct adjustment? Pace et al., 2009
- large p, small n asymptotics: time series, genetics

... questions

- compatibility — Yi, CMS talk
- can composite likelihood be used for modeling when no multivariate distribution exists that is compatible with margins?
- e.g. extreme values, survival data — Parner, 2001
- Hammersley-Clifford theorem for conditional distributions
- analogue for marginal distributions?
- Does theory of multivariate copulas help in understanding this?
- Example: pair specific parameters

$$CL(\omega) = \prod_i \prod_{r<s} f(y_{ir}, y_{is}; \omega_{rs}), \quad \theta = A\omega_{rs}$$

Molenberghs & Verbeke, 2005; Fieuws et al, 2007

... questions

- How do we ensure identifiability of parameters? Yi, CMS talk
- Relationship to modelling via GEE?
- how to investigate robustness systematically?
- how to make use of objective function
- design of composite likelihoods Lindsay, Yi & Sun, 2009
- can we really think beyond means and covariances in multivariate settings?

References

- Varin, C., Reid, N. and Firth, D. (2009). An overview of composite likelihood methods.
- Varin, C. (2008) *Adv. Stat. Anal.* **92**, 1–28. www.dst.unive.it/~sammy
- Lindsay, B. (1988) *Contemp. Math.* **80** 221–240
- 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

Special issue of Statistica Sinica 2010
http://www3.stat.sinica.edu.tw/statistica/