Composite Likelihood: Some Biomedical Applications

Kung-Yee Liang and Chongzhi Di
Department of Biostatistics
Johns Hopkins University

Workshop on Composite Likelihood Methods
April 15-17, 2008, Warwick, England

Outline

- Challenges associated with likelihood inference
- Alternative (likelihood) approaches
- Biomedical applications of composite likelihood
  - Familial aggregation
  - Missing data in regression
  - Case-control study with ordinal responses
- Discussion

Likelihood Inference

Likelihood inference has been successful in a variety of scientific fields

- LOD score method for genetic linkage
  - BRCA1 for breast cancer

  Hall et al. (1990) Science
- Poisson regression for environmental health
  - Fine air particle (PM_{10}) for increased mortality in total cause and in cardiovascular and respiratory causes

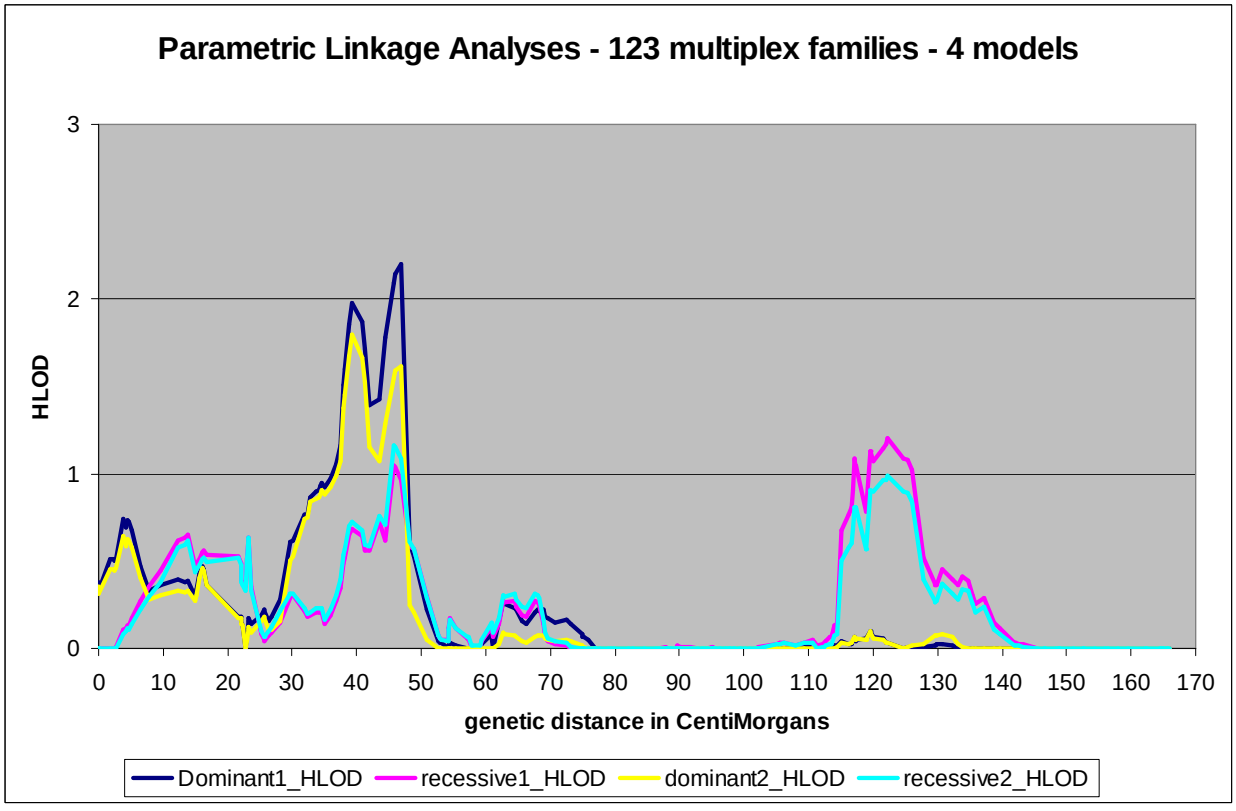
  Samet et al. (2000) NEJM
- ML image reconstruction estimate for nuclear medicine
  - Diagnoses for myocardial infarction and cancers

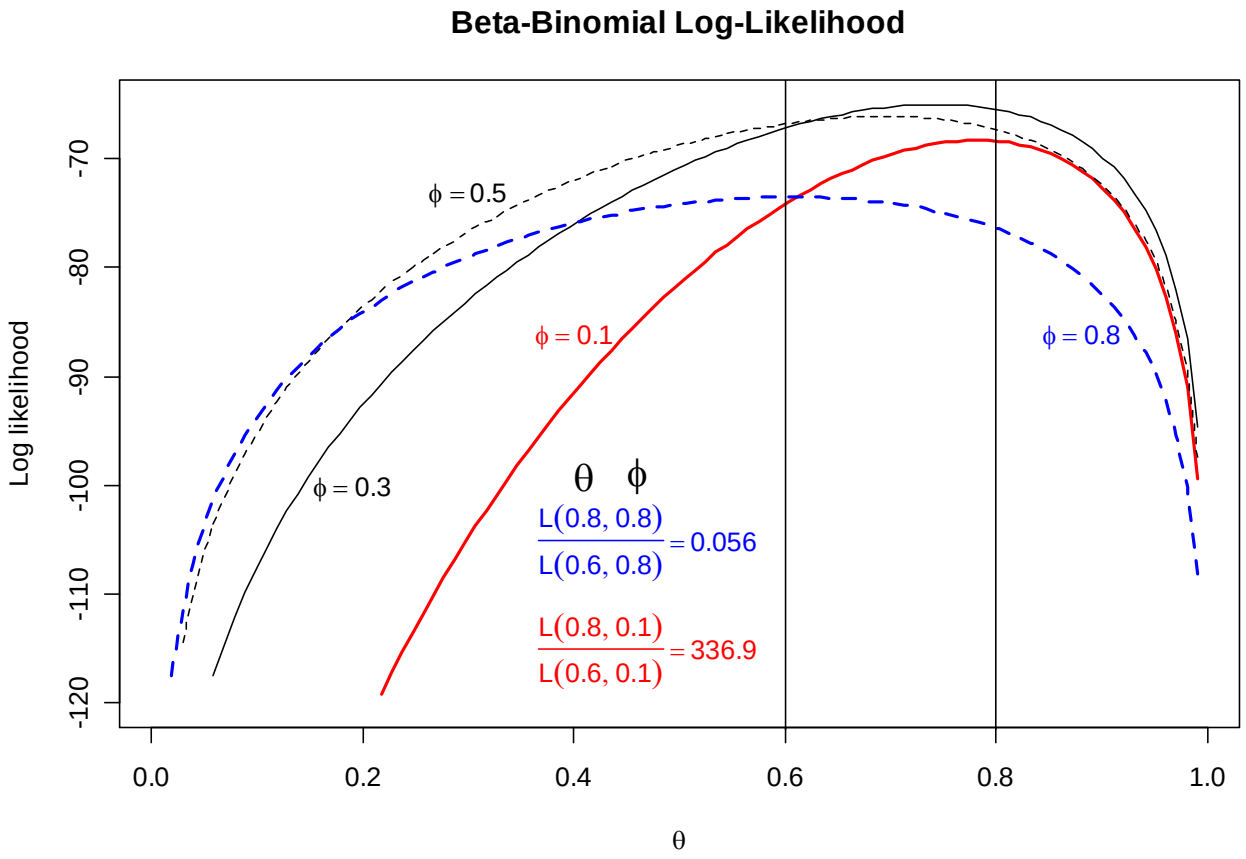
Challenges for Likelihood Inference

- In the absence of sufficient substantive knowledge, likelihood function maybe difficult to fully specify
  - Genetic linkage for complex traits
  - Genome-wide association with thousands of SNPs
  - Gene expression data for tumor cells
- There is computational issue as well for high-dimensional observations
  - High throughput data

Challenges for Likelihood Inference (con't)

- Impacts of nuisance parameters
  - Inconsistency of MLE with many nuisance parameters (Neyman-Scott problem)
  - Different scientific conclusions with different nuisance parameter values
  - Ill-behaved likelihood function
    - Asymptotic approximation not ready





Beta-Binomial Example

- Sensitivity of LR to nuisance values

  $L(\theta = .8, \varphi = 0.8)/ L(\theta = .6, \varphi = 0.8) = 0.056$

  $L(\theta = .8, \varphi = 0.1)/ L(\theta = .6, \varphi = 0.1) = 336.9$

- Asymptotic approximation

Model	Bias	s.d.	s.e.	Lower	Upper
$\varphi_T = \varphi_C$	.363	.488	.473	10.8	0.1
$\varphi_T \neq \varphi_C$	.091	.517	.414	14.2	3.2

Alternative Likelihood Approaches

- Conditional/partial likelihood
  - Useful for eliminating nuisance parameters
  - Limited to particular families of distributions
- Marginal likelihood
  - Particularly useful for variance components
  - Lack of systematic treatment
- Quasi likelihood
  - Only first two moments needed
  - Pathway may not be unique

Alternative Likelihood Approaches (con't)

- Pseudo likelihood II
  - Useful when parameters of interest (θ) and nuisance parameters (φ) are highly intertwined
  - No simple guidance for finding

  Gong & Samaniego (1981) Annals of Statistics $\hat{\varphi}$
- Pseudo likelihood I
  - Focus on scientific questions of interest directly
  - "Cohesiveness" is a challenge
  - A special case (or precursor) of composite likelihood

  Besag (1974) JRSSB
- Empirical likelihood, dual likelihood, etc.

Composite Likelihood

Composition of conditional/marginal likelihoods, which are part of full likelihood components

- Avoiding computational burden
- Making fewer assumptions
  - More robust
- Reducing impacts of nuisance parameters
- Tackling scientific questions of interest more directly
  - Spirit of semi-parametric approaches

Some “Technical” Challenges

- With multiple strata, how to combine contribution from each stratum optimally?
  - Optimum estimating functions
- Asymptotic behavior of MLE’s and LR statistics based on composite likelihoods
  - Characterization of being “information unbiased”
  - Projection method

Some Biomedical Applications

Family case-control study for familial aggregation

- Each case is matched with a control
- Relatives of cases and controls are recruited
- Risk of case relatives (familial risk) is compared with that of control relatives for evidence of familial aggregation

Cohen (1980) Genetic Epidemiology

Liang, Beaty & Cohen (1986) Genetic Epidemiology

Nestadt et al. (2000) Archives of General Psychiatry

Familial Aggregation

Y_{ij}, j = 1,., n_i, affected status of i^{th} case relatives

Y_{ik}, k = n_i +1,., $n_i + m_i$, affected status of i^{th} control relatives

i = 1, …, I

Logit $Pr(Y_{ij} = 1|x_{ij}, \delta_{ij}) = \alpha_i + x_{ij}^t\beta + \theta\delta_{ij}$, j = 1, …, $n_i + m_i$

x: individual covariates

δ = 1(0) if case (control) relative

- θ: primary parameter of interest
- Challenges: how to eliminate nuisance parameters $\{\alpha_i, i = 1, .,K\}$ while accounting for lack of independence among relatives?

Familial Aggregation (con't)

Idea:

1. Adopt the conditional argument for matched designs to case and control relatives in a pair-wise fashion

$\Pr(y_{ij}, y_{ik} | y_{ij} + y_{ik} = t, x_{ij}, \delta_{ij} = 1, x_{ik}, \delta_{ik} = 0) =$

$t = 0,1,2,\ j = 1, \ldots, n_i,\ k = n_i + 1, .., n_i + m_i$

2. Assemble these conditional likelihoods within and across strata together to form the composite likelihood

Liang (1987) Biometrics

Familial Aggregation (con't)

In the absence of covariates (data be summarized in I 2x2 tables), it gives rise to the Mantel-Haenszel estimator with weights $1/(n_i + m_i)$

Composite likelihood methods provide

- A way to extend M-H method to account for additional covariates in logistic regression setting
- Connection between M-H procedure and conditional MLEs by comparing n_i cases with m_i controls simultaneously

Missing Data in Regression

In situations where an individual's chance of missing depends on the outcome value, y, but not on covariates, x, one has

$$f(y|x, \delta = 1) = pr(\delta = 1|y)f(y|x; \beta)/Pr(\delta = 1|x)$$
$$= a(y)\, b(x)\, f(y|x; \beta)$$

$\delta = 1$ if observed and 0 if missing

Challenge: can one make inference on β without specifying the missing mechanism?

Missing Data in Regression (con't)

$$f(y|x, \delta = 1) = a(y)\, b(x)\, f(y|x; \beta)$$

Idea:

1. Consider, with $(z_1, .., z_n)$ the order statistics for $(y_1, .., y_n)$,

   $f(y_1, .., y_n|\delta = 1, x_1, .., x_n, z_1, .., z_n) =$

   $\Pi_i f(y_i|x_i; \beta)/\Sigma \Pi_i f(z_i|x_i; \beta)$

   where Σ is summed over all possible permutation of $\{1, 2,.., n\}$

Missing Data in Regression (con't)

Idea:

2. To reduce computational burden, consider this conditional argument in a pair-wise fashion

$$1/\{1 + R(y_j, x_j; y_k, x_k)\}$$

$$R(y_j, x_j; y_k, x_k) = f(y_i|x_k)f(y_k|x_i)/\{f(y_i|x_k)f(y_k|x_i)\}$$

3. A composite likelihood is formed by putting together $\binom{n}{2}$ such conditional likelihood events

- Applicable to missing covariates as well

Liang & Qin (2000) JRSSB

Case-Control Study with Ordinal Outcomes

It is frequent that individuals diagnosed with the same disease are different in severity, stage, etc.

Questions:

- Can such information be incorporated in analysis in case-control studies?
- Will this lead to more efficient approach?

Ordinal Case-Control Study (con't)

Idea:

1. Consider the adjacent logistic regression model:

$$\log \Pr(Y = j+1)/\Pr(Y = j) = \alpha_j + \beta^t x,\ j = 1,..,C-1$$

- A special case of "stereotype model" by Anderson (1984, JRSSB)

$$\log \Pr(Y = j)/\Pr(Y = 1) = \alpha_j^* + \varphi_j \beta^t x,\ j = 2,..,C$$

$$0 = \varphi_1 \leqslant \varphi_2 \ldots \leqslant \varphi_C$$

with $\varphi_j = j$, $j = 2,..,$ C and $\alpha_j^* = \alpha_1 + .. + \alpha_j$

Ordinal Case-Control Study (con't)

Idea:

2. With retrospective sampling, consider the following conditional likelihood argument (Farewell, 1979, Biometrika)

$$\Pr(Y = j|x, \delta = 1) = \exp(\alpha_j^+ + j\beta^t x)/D$$

$$D = 1 + \Sigma_k \exp(\alpha_k^+ + k\beta^t x)$$

$\delta = 1$ if sampled and $= 0$ otherwise

$$\alpha_j^+ = \alpha_j^* \Pr(\delta = 1|Y = j)/\Pr(\delta = 1|Y = 1)$$

- This gives rise to a composite likelihood for β and $\{\alpha_j^+ , j = 2, .., C\}$

Ordinal Case-Control Study (con't)

Some implications behind this composite likelihood:

- It is important that sampling, while depends on Y, be independent of x

$$\Pr(\delta = 1|Y = j, x) = \Pr(\delta = 1|Y = j)$$

- Intercepts $\{\alpha_j^*, j = 1, .., C\}$ not estimable
- Existing packages for adjacent and stereotype models can be applied for retrospective designs
  - R package "gnm" (Turner and Firth)
  - R package "VGAM" (Thomas W. Yee)

A Genetic Study on Schizophrenia

Schizophrenia is a psychiatric disorder that is

- High in prevalence
- Strong in genetic components (no genes have been found yet though)
- A special case of complex disorders encountering G-G and G-E interactions, genetic heterogeneity, imprinting, etc.

Genetic linkage on chromosome 8 has been reported

Blouin et al. (1998) Nature Genetic

A Genetic Study on Schizophrenia (con't)

Pattern of severity for schizophrenia:

1: Episodic shift

2: Mild deterioration

3: Moderate deterioration

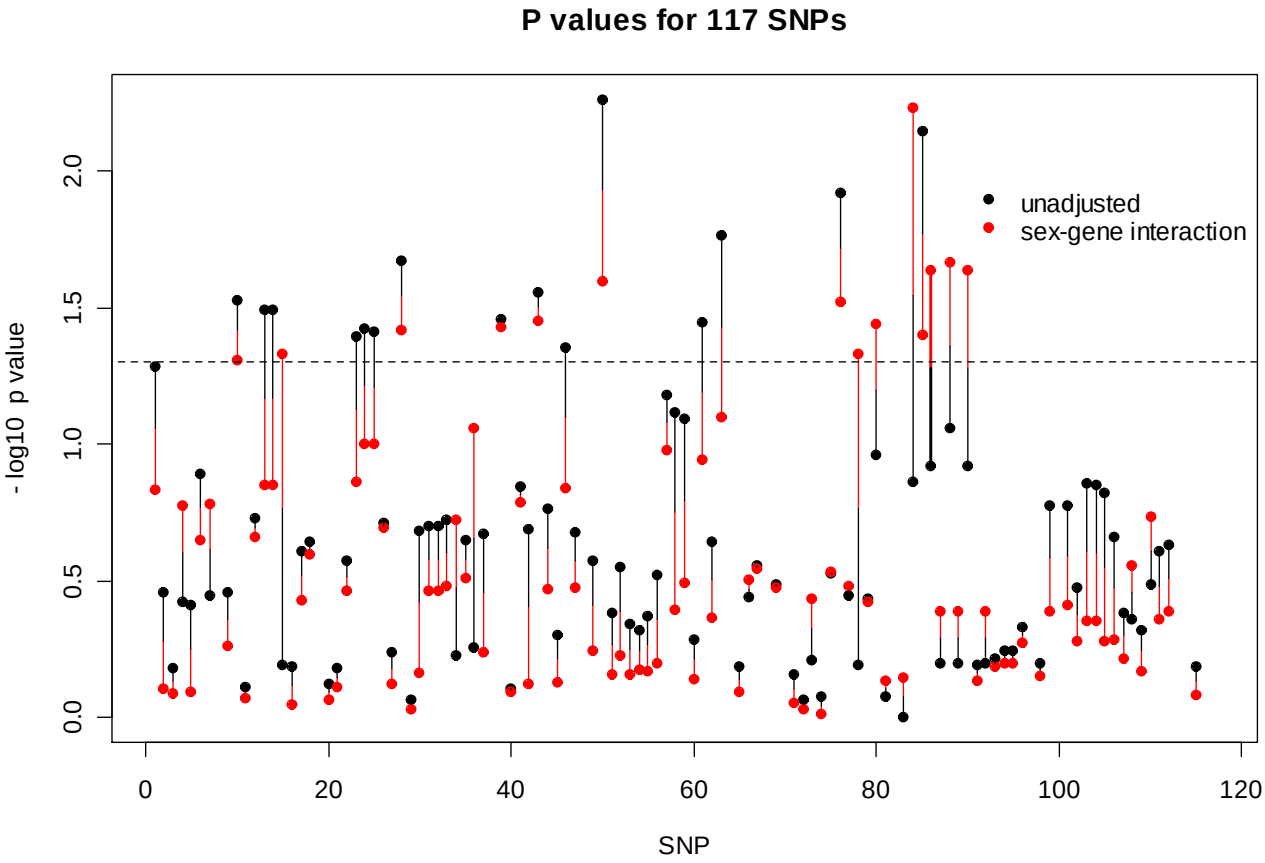
4: Severe deterioration

A Genetic Study on Schizophrenia (con't)

- Frequency table by sex

	Control	Case severity				Total
	0	1	2	3	4	
Male	172	4	51	102	92	421
Female	223	2	23	50	35	333
Total	395	6	74	152	127	754

- 117 SNPs in two genes on Chromosome 8
  DPYSL2 (93 SNPs), PNOC (24 SNPs)



SNP 86: Binary analysis

- Male

G-type	Control	Case
22	112	135
12	48	95 (1.64*)
11	12	19 (1.31)

Female

G-type	Control	Case
22	144	78
12	75	26 (0.64)
11	4	6 (2.77)

- Combined

G-type	Control	Case
22	256	213
12	123	121 (1.18)
11	16	25 (1.88)

SNP 86: Ordinal analysis

- Male

G-type	0	1	2
22	112	86	49
12	48	61 (1.66*)	34 (0.98)
11	12	10 (1.09)	9 (1.58)

Female

G-type	0	1	2
22	144	53	25
12	75	19 (0.69)	7 (0.78)
11	4	3 (2.04)	3 (2.12)

- Combined

G-type	0	1	2
22	256	139	74
12	123	80 (1.20)	41 (0.96)
11	16	13 (1.50)	12 (1.73)

Deviance Tables

Models	Deviance	L.R.	D.F.	P-value
		Binary response		
Gene	1039.31	4.24	2	0.12
Sex	991.99			
G +S	990.25	1.74	2	0.42
G*S	980.66	11.33	4	0.023
		Ordinal response		
Gene	1504.47	5.59	2	0.06
Sex	1462.25			
G + S	1459.45	2.80	2	0.25
G*S	1450.44	11.81	4	0.019

Summary of Results

For SNP 86 (rs6987220),

- It is important that interaction with gender be taken into account
  - Stronger association with risk of schizophrenia among females
  - Recessive with allele 1
- It helps to strengthen finding using ordinal response

Rationale for considering gender:

- 2 to 1 ratio for male vs female cases
- Higher familial risk for female cases
- Gender difference in neuro-development

Summary of Results (con't)

Use of proportional odds models (McCullagh, 1980, JRSSB)

- No need to assign "scores" on ordinal response
- Interpretation of regression coefficient unaffected by "collapsing" adjacent categories
- Application to retrospective sampling less obvious

Discussion

- Composite likelihood provides a useful approach for scientific inference
  - Avoiding undue computational burden
  - Making few assumptions that maybe difficult to verify
  - Reducing non-trivial impacts of nuisance parameters
  - Devoting energy to scientific questions of interest
- With trend of high-dimensional interdependency per subject, this approach and its extension would draw greater attention in statistical community