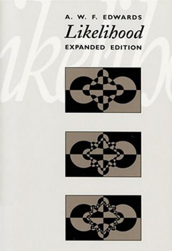
Likelihood inference for complex data

Nancy Reid

Kuwait Foundation Lecture
DPMMS, University of Cambridge
May 5, 2009


A. W. F. EDWARDS
Likelihood
EXPANDED EDITION

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)

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

- observed and expected information:

$$j(\hat{\theta}) = -\ell''(\hat{\theta}; y), \qquad J(\theta) = E_{\theta}\{-\ell''(\theta; y)\}$$

Example: time series studies of air pollution

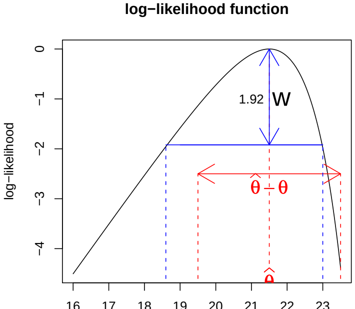
- y_i: number of deaths in Cambridge due to cardio-vascular or respiratory disease on day i
- x_i: 24 hour average of PM_{10} or O_3 in Cambridge on day i, maximum temperature, minimum temperature, dew point, relative humidity, day of the week, ...
- model: Poisson distribution for counts
- $$f(y_i; \theta) = \{\mu_i(\theta)\}^{y_i} \exp\{-\mu_i(\theta)\}$$
- $$\log \mu = \alpha + \psi PM_{10} + S(time, df_1) + S(temp, df_2)$$
- $\theta = (\alpha, \psi, ...)$ with dimension ??
- $S(time, df_1)$ 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

Example: clustered binary data

- latent variable:
  $z_{ir} = x_{ir}'\beta + b_i + \epsilon_{ir}, \quad b_i \sim N(0, \sigma_b^2), \quad \epsilon_{ir} \sim N(0, 1)$
- $r = 1, \ldots, n_i$: observations in a cluster/family/school...
  $i = 1, \ldots, n$ clusters
- random effect b_i introduces correlation between observations in a cluster
- observations: $y_{ir} = 1$ if $z_{ir} > 0$, else 0
- $Pr(y_{ir} = 1 \mid b_i) = \Phi(x_{ir}'\beta + b_i) = p_i \quad \Phi(z) = \int^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$
- likelihood $\theta = (\beta, \sigma_b)$
  $L(\theta; y) = \prod_{i=1}^n \log \int_{-\infty}^{\infty} \prod_{r=1}^{n_i} p_i^{y_{ir}} (1 - p_i)^{1-y_{ir}} \phi(b_i, \sigma_b^2) db_i$
- more general: $z_{ir} = x_{ir}'\beta + w_{ir}' b_i + \epsilon_{ir}$

Inference based on the log-likelihood function

- $\hat{\theta} \dot{\sim} N_d\{\theta, j^{-1}(\hat{\theta})\}$ $\qquad j(\hat{\theta}) = -\ell''(\hat{\theta}; y)$
- $\sqrt{n}(\hat{\theta} - \theta) j^{1/2}(\hat{\theta}) \xrightarrow{\mathcal{L}} N_d(0, I_d)$
- "θ is estimated to be 21.5 (95% CI 19.5 − 23.5)"
- ${}_{19.5}21.5_{23.5}$ $\qquad \hat{\theta} \pm 2\hat{\sigma}$
- $w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \dot{\sim} \chi^2_d$
- "likelihood based CI for θ with confidence level 95% is (18.6, 23.0)" $\qquad {}_{18.6}21.5_{23.0}$



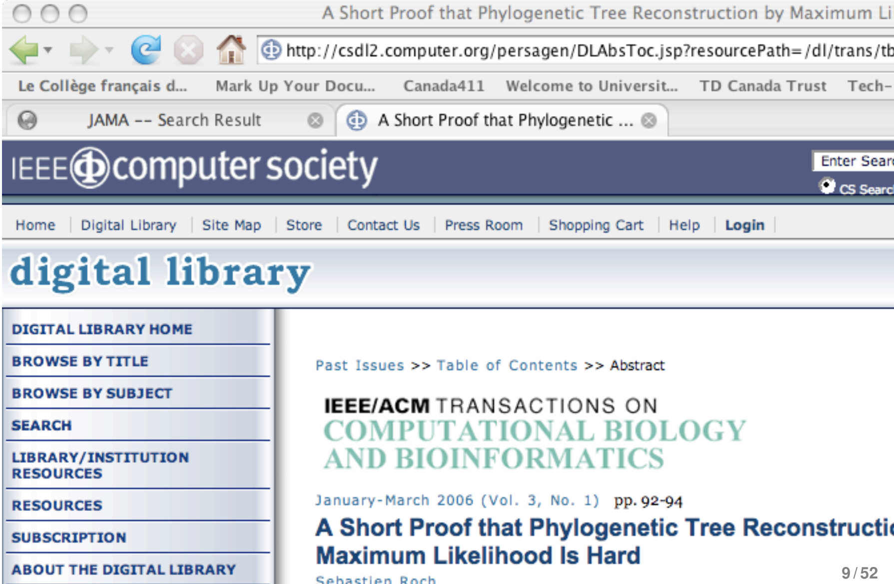
Bayesian inference

- treat θ as a random variable, with a probability distribution and density $\pi(\theta)$
- model interpreted as conditional distribution of y, given θ
- inference for θ based on posterior distribution

$$\pi(\theta \mid y) = \frac{\exp \ell(\theta; y)\pi(\theta)}{\int \exp \ell(\theta; y)\pi(\theta) d\theta}$$

- "θ is estimated to be 21.5, and with 95% probability, θ is between 18.6 and 23.0"
- "using a flat prior density for θ"

Widely used

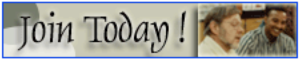

A Short Proof that Phylogenetic Tree Reconstruction by Maximum Li
http://csdl2.computer.org/persagen/DLAbsToc.jsp?resourcePath=/dl/trans/tb
Le Collège français d...
Mark Up Your Docu...
Canada411
Welcome to Universit...
TD Canada Trust
Tech-
JAMA -- Search Result
A Short Proof that Phylogenetic ...
IEEE computer society
Enter Sear
CS Searc
Home | Digital Library | Site Map | Store | Contact Us | Press Room | Shopping Cart | Help | Login
digital library
DIGITAL LIBRARY HOME
BROWSE BY TITLE
BROWSE BY SUBJECT
SEARCH
LIBRARY/INSTITUTION RESOURCES
RESOURCES
SUBSCRIPTION
ABOUT THE DIGITAL LIBRARY
Past Issues >> Table of Contents >> Abstract
IEEE/ACM TRANSACTIONS ON
COMPUTATIONAL BIOLOGY
AND BIOINFORMATICS
January-March 2006 (Vol. 3, No. 1) pp. 92-94
A Short Proof that Phylogenetic Tree Reconstructi
Maximum Likelihood Is Hard
Sebastien Roch

Estimating Genotypic Correlations and Their Standard Errors Using Multivariate Restricted Maxi

file:///Users/nancy/talks/waterloo/cropscience.html

Le Collège français d... Mark Up Your Docu... Canada411 Welcome to Universit... TD Canada Trust Tech–

CROP SCIENCE



QUICK SEARCH: [adva
Author: Keyword(s):
Go
Year: Vol: Page:

HOME HELP FEEDBACK SUBSCRIPTIONS ARCHIVE SEARCH TABLE OF CONTENTS

Published online 1 February 2006
Published in Crop Sci 46:642-654 (2006)
DOI: 10.2135/cropsci2005.0191

677 S. Segoe Rd., Madison, WI 53711 USA

CROP BREEDING, GENETICS & CYTOLOGY

Estimating Genotypic Correlations and Their Standard Errors U
Multivariate Restricted Maximum Likelihood Estimation with S
Proc MIXED

James B. Holland[*]

USDA-ARS Plant Science Research Unit, Dep. of Crop Science, Box 7620, North Carolina State University, Raleigh, NC 27

[*] Corresponding author (James_Holland@ncsu.edu)

Plant breeders traditionally have estimated genotypic and phenotypic correlations between traits using the
moments on the basis of a multivariate analysis of variance (MANOVA). Drawbacks of using the method
moments to estimate variance and covariance components include the possibility of obtaining estimates o

The Review of Financial Studies

Maximum Likelihood Estimation of Latent Affine Processes

David S. Bates
University of Iowa

This article develops a direct filtration-based maximum likelihood methodology for estimating the parameters and realizations of latent affine processes. Filtration is conducted in the transform space of characteristic functions, using a version of Bayes' rule for recursively updating the joint characteristic function of latent variables and the data conditional upon past data. An application to daily stock market returns over 1953–1996 reveals substantial divergences from estimates based on the Efficient Methods of Moments (EMM) methodology; in particular, more substantial and time-varying jump risk. The implications for pricing stock index options are examined.

IEEE Transactions on Information Theory

2062 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 52, NO. 5, MAY 2006

Single-Symbol Maximum Likelihood Decodable Linear STBCs

Md. Zafar Ali Khan, *Member, IEEE,* and B. Sundar Rajan, *Senior Member, IEEE*

***Abstract*—Space–time block codes (STBCs) from orthogonal designs (ODs) and coordinate interleaved orthogonal designs (CIOD) have been attracting wider attention due to their amenability for fast (single-symbol) maximum-likelihood (ML) decoding, and full-rate with full-rank over quasi-static fading channels. However, these codes are instances of single-symbol decodable codes and it is natural to ask, if there exist codes other than STBCs form ODs and CIODs that allow single-symbol decoding? In this paper, the above question is answered in the affirmative by characterizing all linear STBCs, that allow single-symbol ML decoding (not necessarily full-diversity) over quasi-static fading channels-calling them single-symbol decodable designs (SDD). The class SDD includes ODs and CIODs as proper subclasses. Further, among the SDD, a class of those that offer full-diversity, called Full-rank SDD (FSDD) are characterized and classified. We then concentrate on square designs and derive the maximal rate for square FSDDs using a constructional proof. It follows that 1) except for $N = 2$, square complex ODs are not maximal rate and 2) a rate one square FSDD exist only for two and four transmit antennas. For nonsquare designs, generalized coordinate-interleaved orthogonal designs (a superset of CIODs) are presented and analyzed. Finally, for rapid-fading channels an equivalent matrix channel representation is developed, which allows the results of quasi-static fading channels to be applied to rapid-fading channels. Using this representation we show that for rapid-fading channels the rate of single-symbol decodable STBCs are independent of the number of transmit antennas and inversely proportional to the**

difference between coded modulation [used for single-input single-output (SISO), single-iutput multiple-output (SIMO)] and space–time codes is that in coded modulation the coding is in time only while in space–time codes the coding is in both space and time and hence the name. STC can be thought of as a signal design problem at the transmitter to realize the capacity benefits of MIMO systems [1], [2], though, several developments toward STC were presented in [3]–[7] which combine transmit and receive diversity, much prior to the results on capacity. Formally, a thorough treatment of STCs was first presented in [8] in the form of trellis codes [space–time trellis codes (STTC)] along with appropriate design and performance criteria.

The decoding complexity of STTC is exponential in bandwidth efficiency and required diversity order. Starting from Alamouti [12], several authors have studied space–time block codes (STBCs) obtained from orthogonal designs (ODs) and their variations that offer fast decoding (single-symbol decoding or double-symbol decoding) over quasi-static fading channels [9]–[27]. But the STBCs from ODs are a class of codes that are amenable to single-symbol decoding. Due to the importance of single-symbol decodable codes, need was felt for rigorous characterization of single-symbol decodable linear

Journal of the American Medical Association

ORIGINAL CONTRIBUTION

Cognitive Behavioral Therapy for Posttraumatic Stress Disorder in Women

A Randomized Controlled Trial

Paula P. Schnurr, PhD
Matthew J. Friedman, MD, PhD
Charles C. Engel, MD, MPH
Edna B. Foa, PhD
M. Tracie Shea, PhD
Bruce K. Chow, MS
Patricia A. Resick, PhD
Veronica Thurston, MBA
Susan M. Orsillo, PhD
Rodney Haug, PhD
Carole Turner, MN
Nancy Bernardy, PhD

Context The prevalence of posttraumatic stress disorder (PTSD) is elevated among women who have served in the military, but no prior study has evaluated treatment for PTSD in this population. Prior research suggests that cognitive behavioral therapy is a particularly effective treatment for PTSD.

Objective To compare prolonged exposure, a type of cognitive behavioral therapy, with present-centered therapy, a supportive intervention, for the treatment of PTSD.

Design, Setting, and Participants A randomized controlled trial of female veterans (n=277) and active-duty personnel (n=7) with PTSD recruited from 9 VA medical centers, 2 VA readjustment counseling centers, and 1 military hospital from August 2002 through October 2005.

Intervention Participants were randomly assigned to receive prolonged exposure (n=141) or present-centered therapy (n=143), delivered according to standard protocols in 10 weekly 90-minute sessions.

Main Outcome Measures Posttraumatic stress disorder symptom severity was the primary outcome. Comorbid symptoms, functioning, and quality of life were secondary outcomes.

Physical Review D

PHYSICAL REVIEW D **73,** 015013 (2006)

Multidimensional mSUGRA likelihood maps

B. C. Allanach

DAMTP, CMS, Wilberforce Road, Cambridge, CB3 0WA, United Kingdom

C. G. Lester

Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE, United Kingdom

(Received 18 November 2005; published 25 January 2006)

We calculate the likelihood map in the full 7-dimensional parameter space of the minimal symmetric standard model assuming universal boundary conditions on the supersymmetry breakin Simultaneous variations of m_0, A_0, $M_{1/2}$, $\tan\beta$, m_t, m_b and $\alpha_s(M_Z)$ are applied using a Marko Monte Carlo algorithm. We use measurements of $b \rightarrow s\gamma$, $(g-2)_\mu$ and $\Omega_{DM}h^2$ in order to const model. We present likelihood distributions for some of the sparticle masses, for the branching $B_s^0 \rightarrow \mu^+\mu^-$ and for $m_{\tilde{\tau}} - m_{\chi_1^0}$. An upper limit of 2×10^{-8} on this branching ratio might be achi the Tevatron, and would rule out 29% of the currently allowed likelihood. If one allows for non-t neutralino components of dark matter, this fraction becomes 35%. The mass ordering allows the im cascade decay $\tilde{q}_L \rightarrow \chi_2^0 \rightarrow \tilde{l}_R \rightarrow \chi_1^0$ with a likelihood of $24 \pm 4\%$. The stop-coannihilation re highly disfavored, whereas the light Higgs region is marginally disfavored.

US Patent Office

US007058142B2

(12) **United States Patent**
Coene et al.

(10) **Patent No.: US 7,058,14**
(45) **Date of Patent: Jun. 6**

(54) **GENERATION OF AMPLITUDE LEVELS FOR A PARTIAL RESPONSE MAXIMUM LIKELIHOOD (PRML) BIT DETECTOR**

(75) Inventors: **Willem M.J. Coene**, Eindhoven (NL); **Renatus J. Van Der Vleuten**, Eindhoven (NL)

(73) Assignee: **Koninklijke Philips Electronics N.V.**, Eindhoven (NL)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **10/403,544**

(56) **References Cited**

U.S. PATENT DOCUMENTS

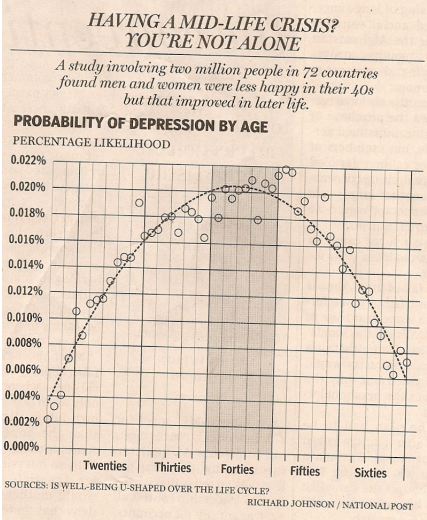
5,113,400	A	5/1992	Gould et al. ....
5,588,011	A	12/1996	Riggle ....
5,666,370	A	9/1997	Ganesan et al. ....
5,764,608	A	6/1998	Satomura ....
5,774,470	A	6/1998	Nishiya et al. ....
6,092,230	A	7/2000	Wood et al. ....
6,278,748	B1	8/2001	Fu et al. ....
6,288,992	B1	9/2001	Okumura et al. ....

Primary Examiner—Pankaj Kumar
(74) *Attorney, Agent, or Firm*—Michael E. Belk

(57) **ABSTRACT**

An apparatus for deriving amplitude values from information signal, which amplitude values can be

In the News



National Post, Toronto, Jan 30 2008

emerged as the more likely cause of his paralysis, they report in the 1 November *Journal of Medical Biography*.

Did FDR Have Guillain-Barré?

A new analysis of Franklin Delano Roosevelt's symptoms suggests he might not have been stricken by polio but by Guillain-Barré syndrome.

In 1921, at the beginning of his political career, Roosevelt became feverish and developed paralysis, which started in his legs and moved up to his neck. Although he recovered partially, he remained permanently wheelchair-bound.

Immunological pediatrician Armond Goldman of the University of Texas Medical Branch in Galveston now says FDR's symptoms are more concordant with Guillain-Barré syndrome, a bacterially induced autoimmune disease. For example,

"The result is interesting both historically and neurologically," says neurologist Deborah Green of the University of Hawaii School of Medicine at Manoa. FDR's misdiagnosis—if such it was—may have changed the course of history, because his affliction gave great momentum to efforts to develop a polio vaccine. But Green notes that "there's no way to prove [a misdiagnosis] without testing the spinal cord fluid." Neurologist H. Royden Jones of Harvard Medical School in Boston adds that the researchers could be wrong in assuming that "Guillain-Barré is the same now as it was back then."

Getting Into a

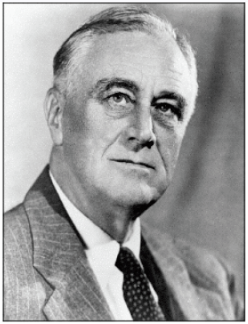


Figure 1. Photograph of President Franklin Delano Roosevelt taken in 1944 by Leon A Perskie. (By permission of Beatrice Perskie Foxman, Silver Springs, Maryland, USA.)

"What was the cause of Franklin Delano Roosevelt's paralytic illness?" Goldman, et al. *J Medical Biography* 2003

Table 2. Diagnostic probabilities of eight key symptoms in Roosevelt's paralytic illness appearing in Guillain–Bar
poliomyelitis, tested by Bayesian analysis

FDR's case	*GBS (prior probability 0.51)*		*Poliomyelitis (pri*
	Symptom probability	*Posterior probability*	*Symptom probability*
Paralysis ascends for 10–13 days	0.70	**0.36**	0.02
Facial paralysis	0.50	**0.26**	0.02
Bladder/bowel dysfunction for 14 days	0.50	**0.26**	0.05
Numbness/dysaesthesia	0.50	**0.26**	<0.01
No meningismus	0.99	**0.50**	0.10
Fever	<0.01	<0.01	0.90
Descending recovery from paralysis	0.70	**0.36**	0.02
Permanent paralysis	0.15	0.08	0.50

The derivation of the estimates of prior probabilities (relative frequencies of the diseases in FDR's age range
probabilities (the chance that a clinical feature occurred in a disease) of poliomyelitis and GBS is given in the
considerations''. Posterior probabilities (the probability that FDR's symptoms were due to a disease) are the
symptom probabilities. Greater posterior probabilities are in bold type.

Variations on a theme

- partial likelihood, Cox, 1972; pseudo-likelihood Besag, 1974, quasi-likelihood Nelder & Wedderburn, 1974
- model part of the data; ignore the other part
- composite likelihood Lindsay, 1988
- profile (concentrated), marginal, conditional, modified profile likelihood
- eliminating nuisance parameters: $\theta = (\psi, \lambda)$
- prequential, predictive likelihood Dawid, 1984; Butler, 1986
- emphasis on predictive performance
- empirical, weighted, robust, bootstrap likelihood
- less dependence on the model
- nonparametric likelihood

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: $\{\mathcal{A}_k, k \in K\}$

- Composite Likelihood: Lindsay, 1988

$$CL(\theta; y) = \prod_{k \in K} L_k(\theta; y)^{w_k}$$

- $L_k(\theta; y) = f(\{y_r \in \mathcal{A}_k\}; \theta)$ likelihood for an event
- $\{w_k, k \in K\}$ a set of weights

Examples

- Composite Conditional Likelihood: Besag, 1974

$$CCL(\theta; y) = \prod_{s \in \mathcal{S}} f_{s|s^c}(y_s \mid y_{s^c}), \quad \mathcal{S} \text{ set of indices}$$

  and variants by modifying events
- Composite Marginal Likelihood:

$$CML(\theta; y) = \prod_{s \in \mathcal{S}} f_s(y_s; \theta)^{w_s},$$

- Independence Likelihood: $\prod_{r=1}^{p} f_1(y_r; \theta)$
- Pairwise Likelihood: $\prod_{r=1}^{p-1} \prod_{s=r+1}^{p} f_2(y_r, y_s; \theta)$
- tripletwise likelihood, ...
- pairwise differences: $\prod_{r=1}^{p-1} \prod_{s=r+1}^{p} f(y_r - y_s; \theta)$
- and even mixtures of *CCL* and *CML*

Derived quantities

- log composite likelihood: $c\ell(\theta; y) = \log CL(\theta; y)$
- score function: $U(\theta; y) = \nabla_\theta c\ell(\theta; y) = \sum_{s\in\mathcal{S}} w_s U_s(\theta; y)$
  $E\{U(\theta; Y)\} = 0$
- 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):

$$G(\theta) = H(\theta)J(\theta)^{-1}H(\theta)$$

- $J \neq H$

Inference

- Sample: $Y_1, \ldots, Y_n$, i.i.d., $CL(\theta; y) = \prod_{i=1}^{n} CL(\theta; y_i)$
- $$\sqrt{n}(\hat{\theta}_{CL} - \theta) \dot{\sim} N\{0, G^{-1}(\theta)\} \qquad G(\theta) = H(\theta)J(\theta)^{-1}H(\theta)$$
- $w(\theta) = 2\{c\ell(\hat{\theta}_{CL}) - c\ell(\theta)\} \dot{\sim} \sum_{a=1}^{d} \mu_a Z_a^2 \quad Z_a \sim N(0, 1)$
- $\mu_1, \ldots, \mu_d$ 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

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

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

- these criteria used for model averaging Hjort and Claeskens, 2008
- or for selection of tuning parameters Gao and Song, 2009

Example: symmetric normal

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

$$
\begin{aligned}
c\ell(\rho; y_1, \ldots, y_n) &= -\frac{np(p-1)}{4}\log(1-\rho^2) - \frac{p-1+\rho}{2(1-\rho^2)}SS_w \\
&\quad - \frac{(p-1)(1-\rho)}{2(1-\rho^2)}\frac{SS_b}{p}
\end{aligned}
$$

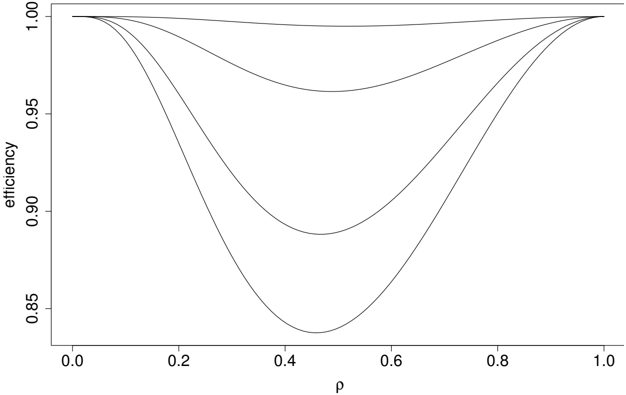
$$
SS_w = \sum_{i=1}^{n}\sum_{s=1}^{p}(y_{is} - \bar{y}_{i.})^2, \quad SS_b = \sum_{i=1}^{n} y_{i.}^2
$$

$$
\begin{aligned}
\ell(\rho; y_1, \ldots, y_n) &= -\frac{n(p-1)}{2}\log(1-\rho) - \frac{n}{2}\log\{1+(p-1)\rho\} \\
&\quad - \frac{1}{2(1-\rho)}SS_w - \frac{1}{2\{1+(p-1)\rho\}}\frac{SS_b}{p}
\end{aligned}
$$

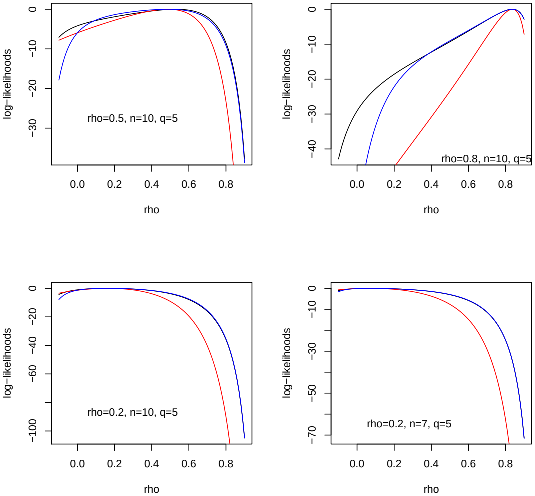
... symmetric normal

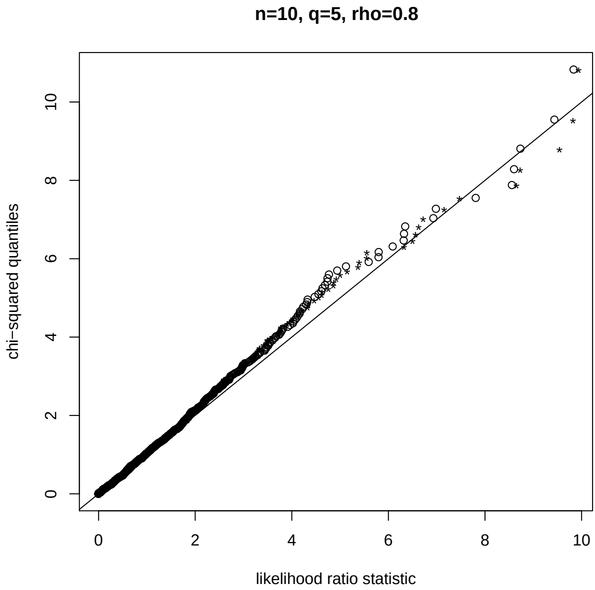
$$\frac{a.var(\hat{\rho}_{CL})}{a.var(\hat{\rho}\,)}, \quad p = 3, 5, 8, 10$$

(Cox & Reid, 2004)



Likelihood ratio test




n=10, q=5, rho=0.8
chi−squared quantiles
likelihood ratio statistic
0
2
4
6
8
10
0
2
4
6
8
10

∗ – pairwise

... symmetric normal +

- $Y_i \sim N(\mu \underline{1}, \sigma^2 R) \qquad R_{st} = \rho$
- $\hat{\mu} = \hat{\mu}_{CL}, \quad \hat{\sigma}^2 = \hat{\sigma}^2_{CL}, \quad \hat{\rho} = \hat{\rho}_{CL}$
- $G(\theta) = H(\theta)J(\theta)^{-1}H(\theta) = J(\theta)$
- pairwise likelihood is fully efficient
- also true for $Y_i \sim N(\mu, \Sigma)$
  (Mardia, Hughes, Taylor 2007; Jin 2009)

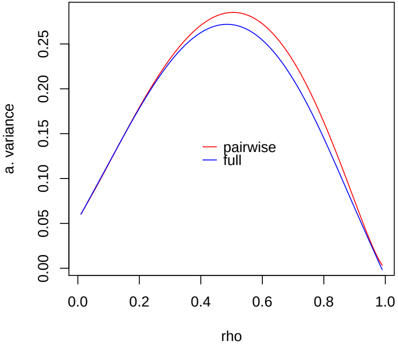
Example: dichotomized MV Normal

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

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

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

$$\text{var}(T) = p^4 (p_{1111} - 2p_{111} + 2p_{11} - p_{11}^2 + \frac{1}{4}) + p^3(-6p_{1111} \ldots) + p^2(\ldots) + p(\ldots)$$



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

Example: clustered binary data

- likelihood

$$L(\beta, \sigma_b) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \prod_{r=1}^{m_i} \Phi(x_{ir}'\beta + b_i)^{y_{ir}} \{1 - \Phi(x_{ir}'\beta + b_i)\}^{1-y_{ir}} \phi(b_i, \sigma_b^2) db_i$$

- pairwise likelihood

$$CL(\beta, \sigma_b) = \prod_{i=1}^{n} \prod_{r<s} P_{11}^{y_{ir}y_{is}} P_{10}^{y_{ir}(1-y_{is})} P_{01}^{(1-y_{ir})y_{is}} P_{00}^{(1-y_{ir})(1-y_{is})}$$

- each $Pr(y_{ir} = j, y_{is} = k)$ evaluated using $\Phi_2(\cdot, \cdot; \rho_{irs})$

(Renard et al., 2004)

... multi-level probit Renard et al. 2004

- computational effort doesn't increase with the number of random effects
- pairwise likelihood numerically stable
- efficiency losses, relative to maximum likelihood, of about 20% for estimation of β
- somewhat larger for estimation of σ_b^2

... Example

D. Renard et al. / Computational Statistics & Data Analysis 44 (2004) 649–667 663

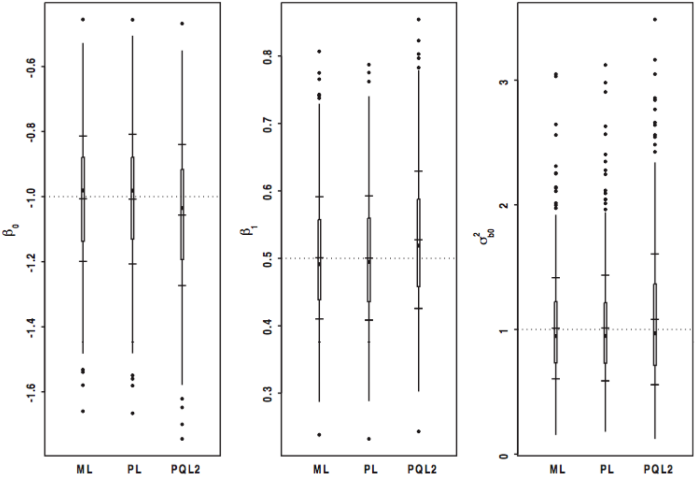


Fig. 5. Boxplots of ML, PL and PQL2 simulated parameter estimates under Model (10) with random intercept.

Markov chains Hjort and Varin, 2008

- comparison of likelihood

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

- adjoining pairs CML

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

- composite conditional likelihood (= Besag's PL)

$$CCL(\theta; y) = \prod \mathrm{pr}(Y_r = y_r \mid \text{ neighbours } ; \theta)$$

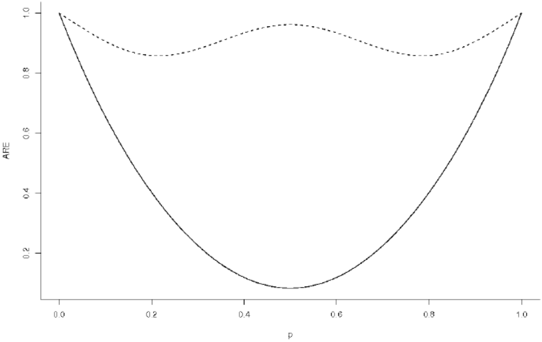
... Markov chain example

- 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 Besag's pseudolikelihood (solid line)



Continuous responses

- Multivariate Normal:
  $Y_i = (Y_{1i}, \ldots, Y_{k_i i}) \sim N\{\beta_0 + \beta_1 x_i, \sigma^2 R_i(\alpha)\}$
  Zhao and Joe, 2005
- pairwise likelihood very efficient, but not $\equiv$ max. lik. ARE
- multivariate longitudinal data; correlated series of observations with random effects
  Fieuws and Verbeke,2006
- correlation of full likelihood and pairwise likelihood estimates of parameters near 1, relative efficiency also near 1 simulations
- pairwise likelihood based on differences within clusters, and connections to within and between block analysis
  Lele and Taper, 2002; Oakes and Ritz, 2000
- and several papers on survival data, often using copulas

CL2			
β_0	β_1	σ^2	ρ
0.998	0.997	1.000	0.913
0.996	0.995	1.000	0.889
0.995	0.996	0.999	0.876
1.000	0.999	1.000	0.884
0.960	0.968	0.987	0.967
0.974	0.970	0.993	0.964
0.978	0.969	0.992	0.928
0.986	0.977	0.993	0.903
0.942	0.958	0.961	0.957
0.944	0.949	0.961	0.952
0.949	0.945	0.966	0.922
0.964	0.939	0.966	0.898
0.924	0.966	0.934	0.943
0.926	0.947	0.937	0.940
0.943	0.932	0.949	0.925
0.982	0.913	0.976	0.919

Binary data

- $Y_r = 1\{Z_r > 0\}$, Z a latent normal r.v.
- generalizations to clustering, longitudinal data: Zhao and Joe 2005, Renard et al 2004
- random effects or multi-level models: Bellio and Varin, 2005; deLeon, 2004
- missing data: Parzen et al, 2007; Yi, Zeng and Cook, 2008
- YZC: not necessary to model the missing data mechanism, uses weighted pairwise likelihood, simulation results promising

... binary data

- questions re choice of weights with clustered data
- comparison of probit and logit
- not clear if marginal parameters and association parameters should be estimated separately
- mixed discrete and continuous data: deLeon and Carriere, 2006; Molenberghs and Verbeke, 2005
- Hybrid pairwise likelihood: GEE for marginal parameters and pairwise likelihood for association parameters: Kuk, 2007
- GEE:

$$\sum_{i=1}^{n} D_i^T V_i^{-1}(y_i - \mu_i) = 0, \quad D_i = \partial\mu_i/\partial\beta$$

Relation to Generalized Estimating Equations

- GEE specifies mean and variance, but not full model
- GEE is fully efficient in multivariate normal model with nonzero correlations
- composite likelihood is fully efficient in a specific multivariate binary model, with a particular dependence model ($\rho_{ir} \neq 0$, ρ_{irs}...all zero)
- composite likelihood seems to be more robust to outliers than GEE
- Qu and Song, 2004 discuss robustness of quadratic inference functions
- composite likelihoods are often easier to maximize
- example: network tomography Liang and Yu, 2003

And more...

- spatial data: multivariate normal, generalized linear models, CML based on differences, CCL and modifications, network tomography, data on a lattice, point processes
- image analysis: Nott and Ryden, 1999
- Rasch model, Bradley-Terry model, ...
- space-time data
- block-based likelihoods for geostatistics
  Caragea and Smith, 2007
- gene mapping (linkage disequilibrium)
  Larribe and Lessard, 2008
- model selection using information criteria based on CL
  Varin and Vidoni, 2005
- improvements of usual CL methods for specific models
- state space models, population dynamics: Andrieu, 2008

Motivation for composite likelihood

- easier to compute:
  - binary data models with random effects, multi-level models (pairwise CML)
  - spatial data: ”near neighbours” CCL – Besag, 1974; Stein, Chi, Welty, 2004
  - sparse networks: Liang and Yu 2003
  - long sequences (large p) in genetics: Fearnhead, 2003; Song, 2007
- access to multivariate distributions:
  - survival data: Parner, 2001; Andersen, 2004, using bivariate copulas
  - multi-type responses, such as continuous/discrete, missing data, extreme values, Oakes and Ritz, 2000; deLeon, 2005; deLeon and Carriere, 2007
- more robust: model marginal (mean/variance) and association (covariance) parameters only

Questions about inference

- Efficiency of composite likelihood estimator:
  - choice of weights: Lindsay, 1988; Kuk and Nott, 2000;
  - assessment by simulation or direct comparison of a. var: Maydeu-Olivares and Joe, 2005
  - comparing two-stage to full pairwise estimation methods: Zhao and Joe, 2005; Kuk, 2007
  - ...
- 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)
- ? Why is CL so efficient (seemingly) ?

Questions about inference

- When Is CML (marginal) preferred to CCL (conditional) ? (always?)
- asymptotic theory: is composite likelihood ratio test preferable to Wald-type test?
- estimation of Godambe information: jackknife, bootstrap, empirical estimates
- estimation of eigenvalues of $(H^{\psi\psi})^{-1}G^{\psi\psi}$
- approximation of distribution of $w(\psi) \dot{\sim} \sum \mu_a Z_a^2$
  - Satterthwaite type? ($f\chi_d^2$): Geys et al, 1999
  - saddlepoint approximation?: Kuonen, 2004
  - bootstrap?
- large p, small n asymptotics: time series, genetics

$p \longrightarrow \infty$

- single long time series
- spatial models (p indexes spatial sites)
- usually assume decaying correlations, so p can play the role of n
- population genetics: 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
- Fearnhead & Donnelly, 2001; McVean et al., 2002; Fearnhead, 2003; Hudson, 2001

... $p \rightarrow \infty$

symmetric normal

$$\text{a.var}(\hat{\rho}_{CL}) = \frac{2}{np(p-1)} \frac{(1-\rho)^2}{(1+\rho^2)^2} c(p^2, \rho^4)$$

$$O(\tfrac{1}{n}) \qquad\qquad O(1)$$

$$n \longrightarrow \infty \qquad\qquad p \longrightarrow \infty$$

dichotomized mv normal:

$$\text{a.var}(\hat{\rho}_{CL}) = \frac{1}{n} \frac{4\pi^2}{p^2} \frac{(1-\rho^2)}{(p-1)^2} \text{var}(T)$$

$$\text{var}(T) = p^4(p_{1111} - 2p_{111} + 2p_{11} - p_{11}^2 + \frac{1}{4}) + \\ p^3(-6p_{1111} \ldots) + p^2(\ldots) + p(\ldots)$$

not consistent if $p \rightarrow \infty, n$ fixed

Questions about modelling

- Is CL useful for modeling when no multivariate distribution exists that is compatible with margins?
- e.g. extreme values, survival data Parner, 2001
- Does theory of multivariate copulas help in understanding this?
- How do we ensure identifiability of parameters?
  – examples of trouble?
- Relationship to modelling via GEE?
- how to investigate robustness systematically?
- E.g. binary data using dichotomized MV Normal
- how to make use of objective function
- can we really think beyond means and covariances in multivariate settings?

.. References

- Firth, Reid and Varin (2010?). An overview of composite likelihood methods. In preparation.
- Special issue of Statistica Sinica (editors Lindsay, Liang and Reid): http://www3.stat.sinica.edu.tw/statistica/

References

- Varin, C. (2008) On composite marginal likelihoods. *Adv. Stat. Anal.* **95**, 1–28 www.dst.unive.it/~sammy
- www.utstat.utoronto.ca/reid/
- Lindsay, B. (1988) *Contemp. Math.* **80** 221–240
- Besag, J. (1974) *JRSS B* **34** 192–236
- Renard, D., Molenberghs, G. and Geys, H. (2004) *Comp. Stat. Data Anal.* **44** 629–667
- Kent, J. (1982) *Biometrika* **69** 19–27
- 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