Topics in Likelihood Inference

STA4508H

Nancy Reid

University of Toronto

February 9, 2022

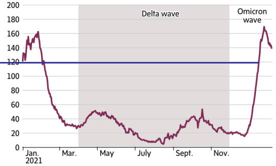
Omicron less lethal than Delta, research finds

Although variant isn't as likely to lead to severe illness, transmissibility has caused daily death counts to eclipse those of previous wave

ANDREA WOO

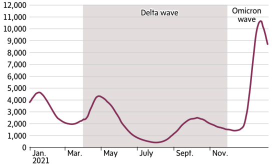
Nurses tend to a COVID-19 patient in the intensive-care unit at a Sarnia, Ont., hospital last month. Canada recorded almost 170 deaths a day during the peak of the Omicron wave in late January, based on a seven-day rolling average. CHRIS YOUNG/THE CANADIAN PRESS

Daily COVID-19 deaths in Canada
Seven-day rolling average



THE GLOBE AND MAIL, SOURCE: GLOBE COVID-19 DATA TRACKER

COVID-19 hospitalization in Canada
Number of people who were hospitalized each day; seven-day rolling average



THE GLOBE AND MAIL, SOURCE: GLOBE COVID-19 DATA TRACKER

People infected with the Omicron variant of SARS-CoV-2 are less likely to die or experience severe outcomes compared with those infected with the Delta variant, emerging research shows.

However, the variant's hyper-transmissibility has led to daily death numbers that have eclipsed those of the previous wave, and a record number of hospitalizations, with Canada's older population still at highest risk.

Federal and provincial governments have begun to signal a turning point in the fight against COVID-19, with some leaders pointing to the variant's relatively less severe outcomes as one reason to lift remaining restrictions within weeks. But experts caution that measures should be eased only when governments have done everything possible to protect the most vulnerable residents.

Samir Sinha, director of geria-

evident among vaccinated and unvaccinated patients and those with or without a prior COVID-19 infection, wrote the authors of the study, which has not yet been peer-reviewed.

The authors note that while such findings are encouraging, the higher transmissibility of Omicron, as well as the fact that prior infection and vaccination are not as effective in fighting off the variant, have led to unprecedented surges of COVID-19 that have

2021, and Jan. 28, 2022.

Of those hospitalized with Delta, 37.5 per cent needed critical care, and 14.2 per cent died. In comparison, 13.1 per cent of those hospitalized with Omicron required critical care, and 6.5 per cent died. The median length of hospital stay was seven days for Delta, and four days for Omicron.

Reviews of 550 cases of people across B.C. admitted to hospital with COVID-19 show that 44 per cent who were hospitalized for

Various ‘types’ of likelihood

1. likelihood, marginal and conditional likelihood, profile likelihood, adjusted profile
2. semi-parametric likelihood, partial likelihood

✓ → GEEs

3. quasi-likelihood, composite likelihood

misspecified models

4. empirical likelihood, penalized likelihood

what to do when assumptions fail

5. likelihood inference in high dimensions
6. simulated likelihood, indirect inference
7. bootstrap likelihood, *h*-likelihood, weighted likelihood, pseudo-likelihood, local likelihood, sieve likelihood

Presentations

Feb 16	Angela: Cox (2013)
	Robert: Barndorff-Nielsen and Cox (1979)
	Shiki: Solomon and Cox (1992)
Feb 23	Hengchao: Rotnitzky et al. (2000)
	Siyue: De Stavola and Cox (2008)
	Manuel: Battey and Cox (2018)
	Ziang: Cox (1975)

Feb 16 in SS 1087; Feb 23 online

typo in exercise 2 for Feb 2 $\exp\{\ell(\theta) - \ell(\hat\theta)\}$ correct version now on web page

exercises Jan 26 has details about report structure

- $y_1, \ldots, y_n$ independent observations $\sim G(\cdot)$, density $g(\cdot)$
- we fit the incorrect model $f(y;\theta)$ $\{ f(y;\theta), \theta \in \Theta \}$

- $y_1, \ldots, y_n$ independent observations $\sim G(\cdot)$, density $g(\cdot)$

of f from g
from f to g

- we fit the incorrect model $f(y; \theta)$
- Kullback-Liebler (KL) divergence between $f(y; \theta)$ and $g(y)$ is defined as

$$KL(\theta) = \int \log \left\{ \frac{g(y)}{f(y; \theta)} \right\} g(y) dy$$

Wikipedia writes $D_{KL}(G||F)$ or more precisely $D_{KL}(P_G||P_F)$

- $KL(\theta) \geq 0$, and $KL(\theta) = 0 \iff f(y; \theta) \equiv g(y)$

- $y_1, \ldots, y_n$ independent observations $\sim G(\cdot)$, density $g(\cdot)$
- we fit the incorrect model $f(y;\theta)$
- Kullback-Liebler (KL) divergence between $f(y;\theta)$ and $g(y)$ is defined as

$$KL(\theta) = \int \log\left\{\frac{g(y)}{f(y;\theta)}\right\} g(y)dy$$

Wikipedia writes $D_{KL}(G||F)$ or more precisely $D_{KL}(P_G||P_F)$

- $KL(\theta) \geq 0$, and $KL(\theta) = 0 \iff f(y;\theta) \equiv g(y)$
- define $\theta^* = \arg\min KL(\theta)$
- $f(y;\theta^*)$ is closest to $\cancel{G}(\cdot)$ $g(\cdot)$ $F(y;\theta^*)$ closest to G in the family $\{f(\cdot;\theta), \theta \in \Theta\}$

… misspecified models

$$KL(\theta) = \int \log\left\{\frac{g(y)}{f(y;\theta)}\right\} g(y)dy$$

- $\theta^* = \arg\min_\theta KL(\theta)$
- $\theta^* = \arg\max_\theta \int \log\{f(y;\theta)\}g(y)dy = \arg\max_\theta \mathrm{E}_G\{\ell(\theta; y)\}$

... misspecified models

$$KL(\theta) = \int \log \left\{ \frac{g(y)}{f(y;\theta)} \right\} g(y) dy$$

- $\theta^* = \arg\min_\theta KL(\theta)$
- $\theta^* = \arg\max_\theta \int \log\{f(y;\theta)\}g(y)dy = \arg\max_\theta \mathrm{E}_G\{\ell(\theta;y)\}$
- leads to a proof that the maximum likelihood estimator converges to θ^*

under some smoothness conditions, etc.

$\ell'(\hat\theta) = 0$ mle under wrong model

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

as hard (early) as proving $\hat{\theta}_n \to \theta$ $= \frac{1}{n}\sum \ell(\theta; y_i) \to E_G\{\ell(\theta; Y)\}$?

$$KL(\theta) = \int \log\left\{\frac{g(y)}{f(y;\theta)}\right\} g(y)dy$$

Knight Math. St.

$\hat{\theta}_n$ max's $\frac{1}{n}\ell(\theta; y)$

i.e. max $\frac{1}{n}\sum_{i=1}^{n} \log f(y_i;\theta)$

- $\theta^* = \arg\min_\theta KL(\theta)$
- $\theta^* = \arg\max_\theta \int \log\{f(y;\theta)\}g(y)dy = \arg\max_\theta \mathrm{E}_G\{\ell(\theta; y)\}$
- leads to a proof that the maximum likelihood estimator converges to θ^*

$\to E_G\{\log f(Y;\theta)\}$

under some smoothness conditions, etc.

- if $g(y) = f(y;\theta_0)$, then $\theta^* = \theta_0$

true density is in the model family

max'd at θ^*

- otherwise θ^* is the 'least false' parameter value

Asymptotic Stat. Van der Vaart

... misspecified models

- Example: true model G is log-normal $\log y \sim N(\mu, \sigma^2)$
- fitted model has density $f(y;\theta) = \frac{1}{\theta}\exp(-\frac{y}{\theta})$

$z \sim N$ (ntbc)

$y = \log z$ is $g(y) = ?$

~~log~~ N

... misspecified models

- Example: true model G is log-normal $\quad \log y \sim N(\mu, \sigma^2)$ $\quad g(y) = ?$
- fitted model has density $\quad f(y;\theta) = \frac{1}{\theta}\exp(-\frac{y}{\theta})$ $\quad \ell(\theta; y) = -\log\theta - \frac{y}{\theta}$
- $E_G\{\ell(\theta; y)\} = -\log\theta - E_G(\frac{y}{\theta})$ $\quad -\frac{\theta}{\theta^2} + E_G(\frac{y}{\theta^2}) = 0 \Rightarrow \theta^* =$
- $\theta^* = E_G(y) = \exp(\mu + \sigma^2/2)$ $\quad \arg\max_\theta E_G\{\ell(\theta; y)\}$

… misspecified models

- Example: true model G is log-normal $\log y \sim N(\mu, \sigma^2)$ $g(y) = ?$
- fitted model has density $f(y;\theta) = \frac{1}{\theta}\exp(-\frac{y}{\theta})$
- $E_G\{\ell(\theta; y)\} = -\log\theta - E_G(\frac{y}{\theta})$
- $\theta^* = E_G(y) = \exp(\mu + \sigma^2/2)$ $\arg\max_\theta E_G\{\ell(\theta; y)\}$
- If we fit $\{f(y;\theta) : \theta > 0\}$ to a sample $y_1, \dots, y_n$ we get $\hat\theta = \bar y$
- WLLN under sampling from $G(\cdot)$, $\bar y \xrightarrow{p} E_G(y) = \theta^*$

θ^* is a 'meaningful' parameter, regardless of the underlying model

... misspecified models

- viewing $\hat{\theta}_n$ as a convenient summary of the data, we can consider properties of likelihood-based inference under the true model g

$\text{var}(\hat{\theta}_n) = ?$ $\quad \widehat{\text{var}}(\hat{\theta}_n) = ?$ $\quad \cdots 2\{\ell(\hat{\theta}) - \ell(\theta)\} \cdots$

Bk

Kent, White, 1982

… misspecified models

- viewing θ as a convenient summary of the data, we can consider properties of likelihood-based inference under the true model g

Kent, White, 1982

- this can be cumbersome: studying robustness to local departures from an assumed model might be more relevant in practice

… misspecified models

- viewing θ as a convenient summary of the data, we can consider properties of likelihood-based inference under the true model g

Kent, White, 1982

- this can be cumbersome: studying robustness to local departures from an assumed model might be more relevant in practice

- composite likelihood is a special type of misspecification   Lindsay, 1988

… misspecified models

- viewing θ as a convenient summary of the data, we can consider properties of likelihood-based inference under the true model g

Kent, White, 1982

- this can be cumbersome: studying robustness to local departures from an assumed model might be more relevant in practice

- composite likelihood is a special type of misspecification — Lindsay, 1988

- another is the framework of generalized estimating equations, with dependence modelled by using a 'working covariance' — Liang & Zeger, 1986

quasi-likelihood

… misspecified models

- viewing θ as a convenient summary of the data, we can consider properties of likelihood-based inference under the true model g

  Kent, White, 1982
- this can be cumbersome: studying robustness to local departures from an assumed model might be more relevant in practice
- composite likelihood is a special type of misspecification — Lindsay, 1988
- another is the framework of generalized estimating equations, with dependence modelled by using a 'working covariance' — Liang & Zeger, 1986
- indirect inference also uses a working (simplified) model that is adjusted using simulations from the true model

  Gouerieroux et al, 1993

Likelihood inference in misspecified models

- maximum likelihood estimate as usual: $(\partial/\partial\theta)\ell(\hat\theta; y) = 0$

$U(\hat\theta; y) = 0$

- consistent for θ^*, the 'least-false' value
- $E_G U(\theta) = E_G(\partial/\partial\theta)\ell(\theta; y) = \int(\partial/\partial\theta)\ell(\theta; y)g(y)dy = 0$, only at θ^*

$E_G U(\theta^*) = 0$

$\frac{\partial}{\partial\theta} E_G \ell(\theta; y)$ is max'd @ ~~least~~ θ^*

$= 0$ at θ^*

- maximum likelihood estimate as usual: $(\partial/\partial\theta)\ell(\hat\theta; y) = 0$

- consistent for θ^*, the 'least-false' value

- $\mathrm{E}_G U(\theta) = \mathrm{E}_G(\partial/\partial\theta)\ell(\theta; y) = \int (\partial/\partial\theta)\ell(\theta; y)g(y)dy = 0$, only at θ^*

- $\mathrm{E}_G(-\partial^2/\partial\theta^2)\ell(\theta; y) = \int (-\partial^2/\partial\theta^2)\ell(\theta; y)g(y)dy \equiv H_G(\theta)$ ← sensitivity $= H_G(\theta)$

- $\mathrm{E}_G\{(\partial/\partial\theta)\ell(\theta; y)\}^2 = \int \{(\partial/\partial\theta)\ell(\theta; y)\}^2 g(y)dy \equiv J_G(\theta) \neq H(\theta)$ ← variance

ntbc

$\cancel{\mathrm{E}_\theta\{\ell'(\theta)\}^2 = -\mathrm{E}_\theta\,\ell''(\theta)}$

$1 = \int f(y;\theta)dy \qquad 0 = \int \partial f_\theta(y;\theta)\, dy = \frac{\partial}{\partial\theta}\int \partial_\theta \log f \cdot \cancel{f}\, g(y) = \mathrm{E}\,\ell'(\theta)$

Likelihood inference in misspecified models

- maximum likelihood estimate as usual: $(\partial/\partial\theta)\ell(\hat\theta;y) = 0$
- consistent for θ^*, the 'least-false' value
- $E_G U(\theta) = E_G(\partial/\partial\theta)\ell(\theta;y) = \int(\partial/\partial\theta)\ell(\theta;y)g(y)dy = 0$, only at θ^*
- $E_G(-\partial^2/\partial\theta^2)\ell(\theta;y) = \int(-\partial^2/\partial\theta^2)\ell(\theta;y)g(y)dy \equiv H(\theta)$
- $E_G\{(\partial/\partial\theta)\ell(\theta;y)\}^2 = \int\{(\partial/\partial\theta)\ell(\theta;y)\}^2 g(y)dy \equiv J(\theta) \neq H(\theta)$
- $(\hat\theta - \theta^*) = H(\theta^*)^{-1}U(\theta^*)\{1 + o_p(1)\}$
- $\hat\theta \,\dot\sim\, N\{\theta^*, \mathcal{G}^{-1}(\theta^*)\}$

$u(\hat\theta) = 0$

$= u(\theta) + (\hat\theta - \theta)u'(\theta) + R_n$

$= H_G(\theta)$

$-u'(\theta)$

n.t.b.c.

$U(\theta) = (\partial/\partial\theta)\ell(\theta)$

$u(\theta^*) \xrightarrow{d} N(0, J(\theta))$

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

Examples

- $y_1, \ldots, y_n$ i.i.d. $\sim G$; we assume $f(y;\theta)$ is $N(\mu, \sigma^2)$
- $\hat{\mu} = \bar{y}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$

$\mu^* = \mu_G, \sigma^{*2} = \sigma_G^2$

- $\partial_\theta \ell(\theta; y) = [\Sigma(y_i - \mu)/\sigma^2, \quad -(n/2\sigma^2) + \Sigma(y_i - \mu)^2/(2\sigma^4)]^{T}$

Examples

- $y_1, \ldots, y_n$ i.i.d. $\sim G$; we assume $f(y;\theta)$ is $N(\mu, \sigma^2)$

- $\hat{\mu} = \bar{y}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$

  $\mu^* = \mu_G, \sigma^{*2} = \sigma_G^2$

- $\partial_\theta \ell(\theta; y) = [\Sigma(y_i - \mu)/\sigma^2, \quad -(n/2\sigma^2) + \Sigma(y_i - \mu)^2/(2\sigma^4)]$

- $H(\theta_G) = n \begin{bmatrix} 1/(\sigma_G^2) & 0 \\ 0 & 1/(2\sigma_G^4) \end{bmatrix} \rightarrow E_G\{-\ell''(\theta)\}$

  $H(\theta) = -E_G \ell''(\theta; y)$

  $J(\theta_G) = n \begin{bmatrix} 1/(\sigma_G^2) & \mu_3/(2\sigma_G^6) \\ \mu_3/(2\sigma_G^6) & (\mu_4 - \sigma_G^4)/(4\sigma_G^8) \end{bmatrix} = \mathrm{var}_G\{\ell'(\theta)\}$

  $J(\theta) = \mathrm{Var}_G \ell'(\theta; y)$

  ntbc

$\mu_3 = E_G(y-\mu)^3 \overset{\cdot}{\underset{N}{=}} 0$

$\mathrm{var}(\hat{\theta}) \doteq \frac{1}{n}(H J^{-1} H)$

$2\sigma^4 \underset{N}{\doteq} \mu_4 = E_G(y-\mu)^4$

Examples

- $y_1, \ldots, y_n$ i.i.d. $\sim G$; we assume $f(y;\theta)$ is $N(\mu, \sigma^2)$

- $\hat\mu = \bar y, \quad \hat\sigma^2 = \frac{1}{n}\sum(y_i - \bar y)^2$

$\mu^* = \mu_G, \sigma^{*2} = \sigma^2_G$

- $\partial_\theta \ell(\theta; y) = [\Sigma(y_i - \mu)/\sigma^2, \quad -(n/2\sigma^2) + \Sigma(y_i - \mu)^2/(2\sigma^4)]$

- $H(\theta_G) = n \begin{bmatrix} 1/(\sigma_G^2) & 0 \\ 0 & 1/(2\sigma_G^4) \end{bmatrix}$

$(H J^{-1} H)^{-1}$

$H(\theta) = -\mathrm{E}_G \ell''(\theta; y)$

$J(\theta) = \mathrm{Var}_G \ell'(\theta; y)$

$$J(\theta_G) = n \begin{bmatrix} 1/(\sigma_G^2) & \mu_3/(2\sigma_G^6) \\ \mu_3/(2\sigma_G^6) & (\mu_4 - \sigma_G^4)/(4\sigma_G^8) \end{bmatrix}$$

ntbc

- $\mathcal{G}(\theta_G) = n^{-1} \begin{bmatrix} \sigma_G^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma_G^4 \end{bmatrix}$

not uncorrelated

… examples

- linear regression model G: $y = X\beta_0 + \epsilon, \quad \epsilon \sim (0, \sigma^2 R)$

  linear regression; correlated errors

- working model $\quad f(y; \beta) = N(X\beta, \sigma^2 I)$

  uncorrelated errors

- $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$

  least squares estimator; mle under normality

- linear regression model G: $y = X\beta_0 + \epsilon, \quad \epsilon \sim (0, \sigma^2 R)$ — linear regression; correlated errors
- working model $f(y;\beta) = N(X\beta, \sigma^2 I)$ — uncorrelated errors
- $\hat\beta = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ — least squares estimator; mle under normality
- $\mathrm{E}_G(\hat\beta) = \beta_0$ — LSE unbiased (consistent)
- $\mathrm{var}_G(\hat\beta) = \sigma^2(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}RX(X^{\mathrm{T}}X)^{-1}$ $\geq \mathrm{var}_F\,\hat\beta$ — sandwich variance

$$(X^TX)^{-1}X^TRX(X^TX)^{-1} \geq (X^TX)^{-1}$$

$B - A$ p.s.d. matrix

... examples

$\rho \; 1$

- linear regression model G: $y = X\beta_0 + \epsilon, \quad \epsilon \sim (0, \sigma^2 R)$

linear regression; correlated errors

- working model $\quad f(y; \beta) = N(X\beta, \sigma^2 I)$

uncorrelated errors

- $\hat\beta = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$

least squares estimator; mle under normality

- $\mathrm{E}_G(\hat\beta) = \beta_0$

LSE unbiased (consistent)

$\begin{pmatrix} 1 & \rho & \rho \\ & \ddots & \\ \rho & & 1 \end{pmatrix}$

- $\mathrm{var}_G(\hat\beta) = \sigma^2(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}RX(X^{\mathrm{T}}X)^{-1}$

sandwich variance

- working model variance is incorrect

$\sigma^2(X^{\mathrm{T}}X)^{-1}$

- single intercept model $\hat\beta = \bar y$, $R_{ij} = \rho$, $\mathrm{var}(\bar y) = (\sigma^2/n)\{1 + \rho(n-1)\}$

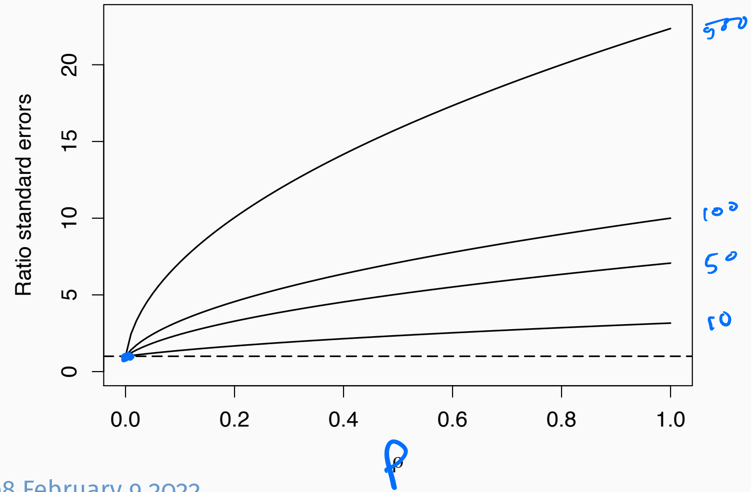
$\frac{\sigma^2}{n}\{ \quad \}$

dependence is a killer

- $y_i \sim N(\beta, \sigma^2), \quad R_{ij} = \rho$

- $\text{var}_G(\hat{\beta}) = (\sigma^2/n)\{1 + \rho(n-1)\}$

$\rho = 0.1, n = 100$ ratio of se's 3.3



$n = 10, 50, 100, 500$

$$\sqrt{1+(n-1)\rho}$$

$$\frac{\sigma^2}{n}\{1+\underline{(n-1)\rho}\}$$

Recap: Composite likelihood

- Vector observation: $Y \sim f(y;\theta)$, $\quad Y \in \mathcal{Y} \subset \mathbb{R}^m$, $\quad \theta \in \mathbb{R}^d$
- Set of events: $\{\mathcal{A}_k, k \in K\}$

$\ell(\theta, \eta; Y)$

$\ell(\theta; y)$

- Composite Log-Likelihood:

Lindsay, 1988

$$c\ell(\theta; y) = \sum_{k \in K} w_k \ell_k(\theta; y)$$

$\ell_k(\theta; y)$ $y \in A_k$

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

$$c\ell_{ind} = \sum_{r=1}^{m} w_r \log f_{Y_r}(y_r; \theta)$$

- Note that θ is assumed to have the same meaning in ℓ_k and in the full model
- This makes it different from a completely mis-specified model

Derived quantities

sample $y = (y_1, \ldots, y_n)$ with joint density $f(y; \theta)$, $y \in \mathbb{R}^m, \theta \in \mathbb{R}^d$

r ind't

score function	$U_{CL}(\theta) = \frac{\partial}{\partial\theta} c\ell(\theta; y) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} c\ell(\theta; y_i)$
maximum composite likelihood estimate	$\hat{\theta}_{CL} = \hat{\theta}_{CL}(y) = \arg\sup_\theta c\ell(\theta; y)$
score equation	$U_{CL}(\hat{\theta}_{CL}) = c\ell'(\hat{\theta}_{CL}) = 0$
composite LRT	$w_{CL}(\theta) = 2\{c\ell(\hat{\theta}_{CL}) - c\ell(\theta)\}$
Godambe information	$G(\theta) = G_n(\theta) = H_n(\theta) J_n^{-1}(\theta) H_n(\theta) = O(n)$

$u'_{cL}(\hat{\theta}_{cL}) = 0$

$O(n)$ $O(\frac{1}{n})$ $O(n)$

Inference

$\theta \in \mathbb{R}^d$

- Sample: $Y_1, \dots, Y_n$, i.i.d., $CL(\theta; \underline{y}) = \prod_{i=1}^n CL(\theta; y_i)$
- $\hat\theta_{CL} - \theta \mathrel{\dot\sim} N\{0, G_n^{-1}(\theta)\}$ $\qquad G_n(\theta) = H(\theta)J(\theta)^{-1}H(\theta)$
- $U(\hat\theta_{CL}) \doteq U(\theta) + (\hat\theta_{CL} - \theta)\partial_\theta U(\theta)$ $\qquad U = U_{CL}$
- $\hat\theta_{CL} - \theta \doteq -\partial_\theta U(\theta)^{-1} U(\theta) \doteq H^{-1}(\theta)U(\theta)$
- $U(\theta) \mathrel{\dot\sim} N\{0, J(\theta)\}$
- $H^{-1}(\theta)U(\theta) \mathrel{\dot\sim} N\{0, H^{-1}(\theta)J(\theta)H^{-\mathrm{T}}(\theta)\}$
- conclude

$$\sqrt{n}(\hat\theta_{CL} - \theta) \mathrel{\dot\sim} N\{0, G^{-1}(\theta)\}$$

$$(\hat\theta_{CL} - \theta)^T G_n(\hat\theta)(\hat\theta_{CL} - \theta) \mathrel{\dot\sim} \chi^2_d$$

... inference

- $w(\theta) = 2\{c\ell(\hat\theta_{CL}) - c\ell(\theta)\} \dot\sim \sum_{a=1}^{d} \mu_a Z_a^2 \quad Z_a \sim N(0,1)$

$\ell(\hat\theta_{CL}) = \ell(\hat\theta)$

$+(\theta-\hat\theta_{CL})\ell'(\hat\theta_{CL}) \to 0$

- $\mu_1, \ldots, \mu_d$ eigenvalues of $J(\theta)H(\theta)^{-1}$

$$c\ell(\hat\theta_{CL}) - c\ell(\theta) \doteq \frac{1}{2}(\hat\theta_{CL}-\theta)^{\mathrm{T}}\{-c\ell''(\hat\theta_{CL})\}(\hat\theta_{CL}-\theta)$$

$+\frac{1}{2}(\theta-\hat\theta)^2\ell''(\hat\theta_{CL})$

$\xrightarrow{d} N(0, G^{-1}) \qquad \xrightarrow{p} H$

- non-central χ^2 limit

- $J(\theta) = \mathrm{var}U(\theta), \qquad H(\theta) = -\mathrm{E}\partial_\theta U(\theta)$

$G = HJ^{-1}H$

- if $J(\theta) = H(\theta)$, $w(\theta) \dot\sim \chi^2_d$

- if $d = 1$, $w(\theta) \dot\sim \mu_1\chi_1^2 = J(\theta)H^{-1}(\theta)\chi_1^2$

H, *J* both scalars

Example: symmetric normal

$CL = \prod_{i=1}^{n} \prod_{r<s} f_2(y_{ir}, y_{is}; \theta)$

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

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

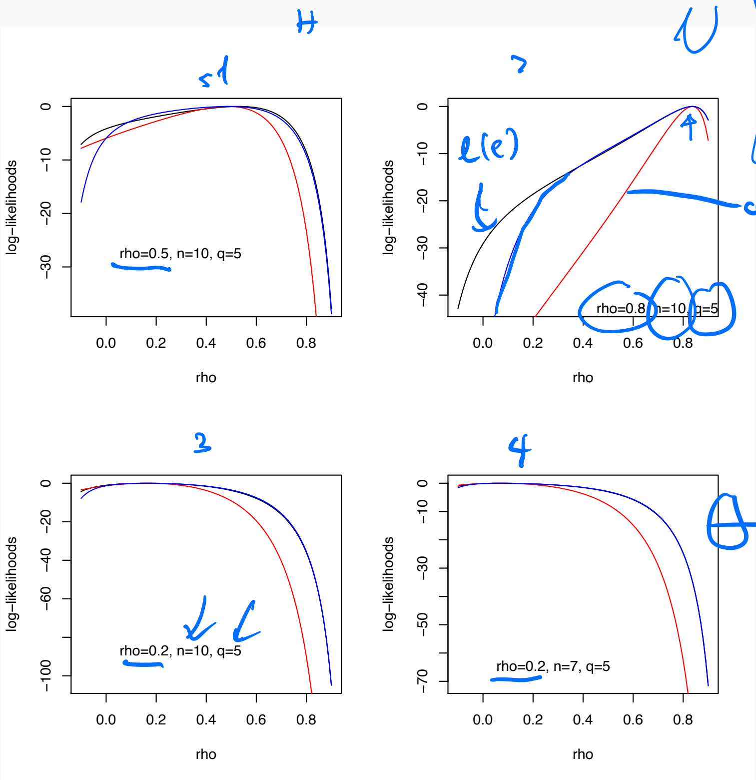
$$
SS_w = \sum_{i=1}^{n} \sum_{s=1}^{m} (y_{is} - \bar{y}_{i.})^2, \quad SS_b = \sum_{i=1}^{n} y_{i.}^2
$$

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 pairwise likelihood

$$
\begin{aligned}
c\ell(\rho; y_1, \ldots, y_n) &= -\frac{nm(m-1)}{4}\log(1-\rho^2) - \frac{m-1+\rho}{2(1-\rho^2)}SS_w \\
&\quad - \frac{(m-1)(1-\rho)}{2(1-\rho^2)}\frac{SS_b}{m} \\
SS_w &= \sum_{i=1}^{n}\sum_{s=1}^{m}(y_{is} - \bar{y}_{i.})^2, \quad SS_b = \sum_{i=1}^{n} y_{i.}^2
\end{aligned}
$$

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



$$N\left[\begin{pmatrix}0\\0\\\vdots\\0\end{pmatrix}, \begin{pmatrix}1 & \rho & \cdots & \rho\\ & \ddots & & \\ & & & \rho\\ & & & 1\end{pmatrix}\right]$$

$$\sim \chi^2_1 \cdot \left(\frac{J}{H}\right)$$

θ

Nuisance parameters $\theta = (\psi, \lambda)$

$\tilde{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$

- constrained estimator: $\tilde{\theta}_\psi = \sup_{\theta=\theta(\psi)} c\ell(\theta; y)$

- $\sqrt{n}(\hat{\psi}_{CL} - \psi) \dot{\sim} N\{0, G^{\psi\psi}(\theta)\}$ $\qquad G(\theta) = H(\theta)J(\theta)^{-1}H(\theta)$

Nuisance parameters $\theta = (\psi, \lambda)$

- constrained estimator: $\tilde{\theta}_\psi = \sup_{\theta=\theta(\psi)} c\ell(\theta; y)$
- $\sqrt{n}(\hat{\psi}_{CL} - \psi) \dot{\sim} N\{0, G^{\psi\psi}(\theta)\}$ $\qquad G(\theta) = H(\theta)J(\theta)^{-1}H(\theta)$
- profile composite log-likelihood test $\quad w(\psi) = 2\{c\ell(\hat{\theta}_{CL}) - c\ell(\tilde{\theta}_\psi)\} \dot{\sim} \sum_{a=1}^{d_0} \mu_a Z_a^2$
- $\mu_1, \ldots, \mu_{d_0}$ are the eigenvalues of $(H^{\psi\psi})^{-1}G^{\psi\psi}$

dim ψ

$$H = -\frac{\partial^2 \ell}{\partial\theta\,\partial\theta^T}$$

$$H^{-1} = \begin{bmatrix} H^{\psi\psi} & H^{\psi\lambda} \\ H^{\lambda\psi} & H^{\lambda\lambda} \end{bmatrix} \qquad \begin{bmatrix} H_{\psi\psi} & H_{\psi\lambda} \\ H_{\lambda\psi} & H_{\lambda\lambda} \end{bmatrix}$$

Nuisance parameters $\theta = (\psi, \lambda)$

- constrained estimator: $\tilde\theta_\psi = \sup_{\theta=\theta(\psi)} c\ell(\theta; y)$

- $\sqrt{n}(\hat\psi_{CL} - \psi) \dot\sim N\{0, G^{\psi\psi}(\theta)\}$ $\qquad$ $G(\theta) = H(\theta)J(\theta)^{-1}H(\theta)$

- profile composite log-likelihood test $\quad w(\psi) = 2\{c\ell(\hat\theta_{CL}) - c\ell(\tilde\theta_\psi)\} \dot\sim \sum_{a=1}^{d_0} \mu_a Z_a^2$

- $\mu_1, \ldots, \mu_{d_0}$ are the eigenvalues of $(H^{\psi\psi})^{-1}G^{\psi\psi}$

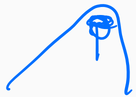
Kent, 1982

- Godambe information needs to be estimated

$$\hat H(\theta) = -\partial^2 c\ell(\hat\theta_{CL})/\partial\theta\partial\theta^{\mathrm{T}}$$

??

$$\hat J(\theta) = n^{-1}\sum_{i=1}^{n} U_{CL}(\theta; y_i)U_{CL}^{\mathrm{T}}(\theta; y_i)$$



latter needs independent samples; can be biased and/or inefficient

Model selection

- Akaike's information criterion

Varin and Vidoni, 2005

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

$H J^{-1} H$

- derivation of AIC for misspecified likelihood leads to

$$TIC = -2c\ell(\hat{\theta}_{CL}) + 2\,\mathrm{tr}\{H(\hat{\theta}_{CL})G^{-1}(\hat{\theta}_{CL})\}$$

Takeuchi information criterion

Model selection

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

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

- derivation of *AIC* for misspecified likelihood leads to

$$TIC = -2c\ell(\hat{\theta}_{CL}) + 2\,\mathrm{tr}\{H(\hat{\theta}_{CL})G^{-1}(\hat{\theta}_{CL})\}$$

Takeuchi information criterion

- Bayesian information criterion — Gao and Song, 2009

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

used for selection of tuning parameters

? CLIC ?

Example: CL with dichotomized MV Normal

$\sim Bin(p_1)$

$\boxed{Y_{ir} = 1\{Z_{ir} > 0\}}$ $\boxed{Z \sim N(0, R)}$ $r = 1, \ldots, m, i = 1, \ldots, n$

$\begin{pmatrix} 1 & & \rho \\ & \ddots & \\ \rho & & 1 \end{pmatrix}$

pairwise

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

$P_{11}^{(\rho)} = \Pr(Z_{ir} > 0, Z_{is} > 0)$

$=$ bivariate N dist of (Z_{ir}, Z_{is})

$\sim \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\}$

Example: CL with dichotomized MV Normal

$$Y_{ir} = 1\{Z_{ir} > 0\} \qquad Z \sim N(0, R) \qquad r = 1, \dots, m; i = 1, \dots, n$$

$\ell(\rho)$ ~ [MVN]

∴ Fisher info.

$P_{......}$

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

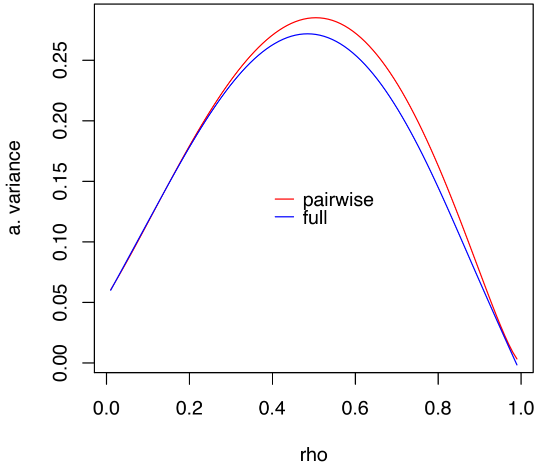
$G^{-1}(\rho)$

$$\text{a.var}(\hat{\rho}_{CL}) = \frac{1}{n} \frac{4\pi^2}{m^2} \frac{(1 - \rho^2)}{(m - 1)^2} \text{var}(T) \qquad T = \sum_{s<r} (2y_{ir}y_{is} - y_{ir} - y_{is})$$

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

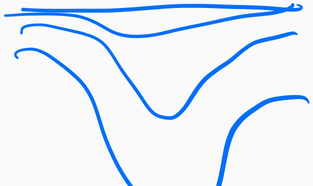
$$p_{1111} = Pr(Z_r > 0, Z_s > 0, Z_t > 0, Z_u > 0)$$

$Pr_{\rho}(4\ Z\text{'s} > 0)$



$q = 3 \quad 5 \quad 10$

20



Numbers incorrect in Cox & Reid 2004 Table 1

ρ	0.02	0.05	0.12	0.20	0.40	0.50
ARE	0.998	0.999	0.995	0.992	0.968	0.953

ρ	0.60	0.70	0.80	0.90	0.95	0.98
ARE	0.938	0.903	0.900	0.874	0.869	0.850

Example: multi-level probit model

- latent variable: $z_{ir} = x_{ir}'\beta + b_i + \epsilon_{ir}$, $\quad \epsilon_{ir} \sim N(0,1)$
- binary observations: $y_{ir} = 1(z_{ir} > 0)$; $\quad r = 1, \ldots m_i$; $i = 1, \ldots n$
- probit model: $Pr(y_{ir} = 1 \mid b_i) = \Phi(x_{ir}'\beta + b_i)$; $\quad b_i \sim N(0, \sigma_b^2)$

z_{ir} $\quad z_{is}$

b_i

Example: multi-level probit model

Renard et al. 2004

- latent variable: $z_{ir} = x_{ir}'\beta + b_i + \epsilon_{ir}, \quad \epsilon_{ir} \sim N(0,1)$
- binary observations: $y_{ir} = 1(z_{ir} > 0); \quad r = 1, \ldots m_i; i = 1, \ldots n$
- probit model: $Pr(y_{ir} = 1 \mid b_i) = \Phi(x_{ir}'\beta + b_i); \quad b_i \sim N(0, \sigma_b^2)$

y

$x_i^T\beta + \epsilon_i$

$y_i \sim N(x_i^T\beta, \sigma^2)$

- 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

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

$w_{ir}^T b_i$

- latent variable: $z_{ir} = x_{ir}'\beta + b_i + \epsilon_{ir}, \quad \epsilon_{ir} \sim N(0,1)$
- binary observations: $y_{ir} = 1(z_{ir} > 0); \quad r = 1, \dots m_i; i = 1, \dots n$
- probit model: $Pr(y_{ir} = 1 \mid b_i) = \Phi(x_{ir}'\beta + b_i); \quad b_i \sim N(0, \sigma_b^2)$

- likelihood

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

ntbc

- pairwise likelihood

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

$x_{ir}^T\beta \quad x_{is}^T\beta$

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

… multi-level probit

- 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

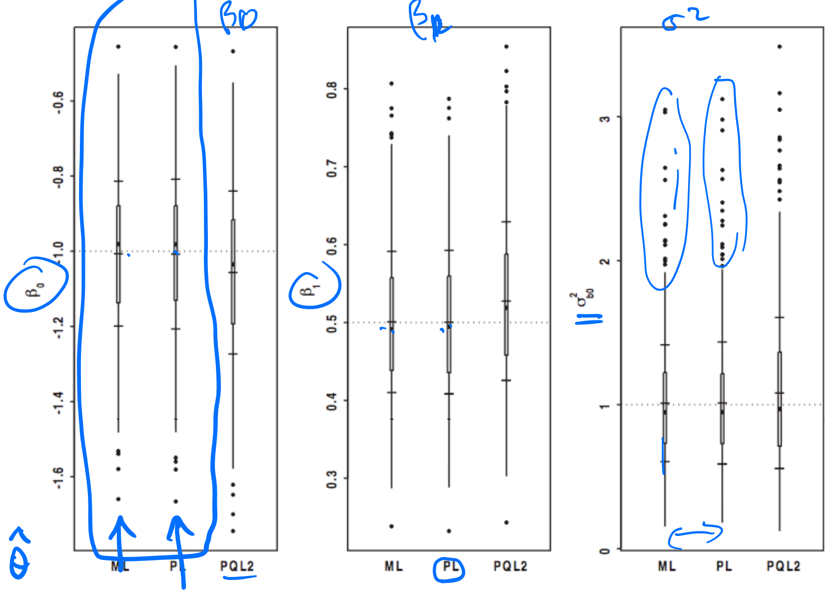


Fig. 5. Boxplots of ML, PL and PQL2 simulated parameter estimates under Model (10) with random intercept.

$\hat{\theta}_{CL}$

penalized quasi-lik.

Example: longitudinal count data

Henderson & Shimakura, 2003

- subjects $i = 1, \dots, n$
- observations counts $y_{ir}, r = 1, \dots m_i$
- model $y_{ir} \sim \text{Poisson}(u_{ir} x_{ir}^T \beta)$
- $u_{i1}, \dots, u_{im_i}$ gamma-distributed random effects

- subjects $i = 1, \ldots, n$
- observations counts $y_{ir}, r = 1, \ldots, m_i$
- model $y_{ir} \sim \text{Poisson}(u_{ir} x_{ir}^T \beta)$
- $u_{i1}, \ldots, u_{im_i}$ gamma-distributed random effects
- but correlated $\text{corr}(u_{ir}, u_{is}) = \rho^{|r-s|}$
- joint density has combinatorial number of terms in m_i; impractical
- weighted pairwise composite likelihood

$$cL_{pair}(\beta) = \prod_{i=1}^{n} \frac{1}{m_i - 1} \prod_{r=1}^{m_i} \prod_{s=r+1}^{m_i} f(y_{ir}, y_{is}; \beta)$$

- weights chosen so that $\mathcal{L}_{pair}$ = full likelihood if $\rho = 0$

$CL(\theta; y)$

$\propto \prod_{k} L_k(\theta; y)^{w_k}$

Example: Varin & Czado 2010

- pain severity scores recorded at four time points → each day

  morning, noon, evening, bed

- 119 patients; varying number of days per patient
- covariates: personal and weather
- response: pain score 0 1 2 3 4 5

Example: Varin & Czado 2010

- pain severity scores recorded at four time points

morning, noon, evening, bed

- 119 patients; varying number of days per patient
- covariates: personal and weather
- response: pain score 0 1 2 3 4 5

- y_{ij} response at time t_{ij} for observation j on subject i, $j = 1, \ldots, m_i$
- y^*_{ij} a latent variable, continuous $y^*_{ij} = x_{ij}^{\mathrm{T}}\beta + u_i + \epsilon_{ij}$
- $y_{ij} = k \Leftrightarrow a_{k-1} < y^*_{ij} < a_k$
- if $u_i \sim N(0, \sigma^2)$ and $\epsilon_{ij} \sim N(0, 1)$

3

α_2 α_3

$$L(\theta; y) = \prod_{i=1}^{n} f(y_{i1}, \ldots, y_{im_i}) =$$

$$\prod_{i=1}^{n} \int_{-\infty}^{\infty} \prod_{j=1}^{m_i} \{\Phi(a_{y_{ij}} - x_{ij}^{\mathrm{T}}\beta - u_i) - \Phi(a_{y_{ij}-1} - x_{ij}^{\mathrm{T}}\beta - u_i)\}\phi(\tfrac{u_i}{\sigma})du_i$$

$\theta = (\underline{a}, \beta, \sigma^2)$

N random effect

... pain severity scores

- y^*_{ij} and $y^*_{ij'}$ have constant correlation $\sigma^2/(\sigma^2+1)$
- points nearer in time might be expected to have higher correlation
- change ϵ_{ij} i.i.d. $N(0,1)$ to $\text{corr}(\epsilon_{ij}, \epsilon_{ij'}) = \exp(\delta|t_{ij} - t_{ij'}|)$

$\tilde{a}_{ij} = a_{ij} - x_{ij}^{\mathrm{T}}\beta/\sqrt{(\sigma^2+1)}$

... pain severity scores

- y^*_{ij} and $y^*_{ij'}$ have constant correlation $\sigma^2/(\sigma^2+1)$
- points nearer in time might be expected to have higher correlation
- change ϵ_{ij} i.i.d. $N(0,1)$ to $\text{corr}(\epsilon_{ij}, \epsilon_{ij'}) = \exp(\delta|t_{ij} - t_{ij'}|)$

$$\tilde{a}_{ij} = a_{ij} - x_{ij}^{\mathrm{T}}\beta/\sqrt{(\sigma^2+1)}$$

α

$$L(\theta; y) = \prod_{i=1}^{n} \int_{\tilde{a}_{y_{i1}-1}}^{\tilde{a}_{y_{i1}}} \cdots \int_{\tilde{a}_{y_{im_i}-1}}^{\tilde{a}_{y_{im_i}}} \phi_{n_i}(z_{i1}, \ldots, z_{im_i}; R_i) dz_{i1} \ldots dz_{im_i}$$

- $$R_{ijj'} = \frac{\sigma^2}{\sigma^2+1} + \frac{e^{-\delta|t_{ij}-t_{ij'}|}}{\sigma^2+1}$$

... pain severity scores

- y_{ij}^* and $y_{ij'}^*$ have constant correlation $\sigma^2/(\sigma^2+1)$
- points nearer in time might be expected to have higher correlation
- change ϵ_{ij} i.i.d. $N(0,1)$ to $\text{corr}(\epsilon_{ij},\epsilon_{ij'}) = \exp(\delta|t_{ij}-t_{ij'}|)$

$\tilde{a}_{ij} = a_{ij} - x_{ij}^{\mathrm{T}}\beta/\sqrt{(\sigma^2+1)}$

- $$L(\theta;y) = \prod_{i=1}^{n}\int_{\tilde{a}_{y_{i1}-1}}^{\tilde{a}_{y_{i1}}}\cdots\int_{\tilde{a}_{y_{im_i}-1}}^{\tilde{a}_{y_{im_i}}}\phi_{n_i}(z_{i1},\ldots,z_{im_i};R_i)dz_{i1}\ldots dz_{im_i}$$

$$R_{ijj'} = \frac{\sigma^2}{\sigma^2+1} + \frac{e^{-\delta|t_{ij}-t_{ij'}|}}{\sigma^2+1}$$

- pairwise log-likelihood:

$$c\ell(\theta;y) = \sum_{i=1}^{n}\sum_{j<j'}^{m_i}\log f_2(y_{ij},y_{ij'};\theta)\boxed{1_{[-q,q]}(t_{ij}-t_{ij'})}$$

weights are 1 or 0, depending on distance between time points

α are cut points

Table 5: Migraine data. Estimates and standard errors from the pairwise likelihood with $q = 12$ for the base model (first two columns) and the best model (last two columns) accordingly to CLIC. The levels of the variable **change** are 1: change from low to high atmospheric pressure, 2: substantially unchanged atmospheric pressure, 3: change from high to low atmospheric pressure. The baseline is "no university degree, no intake of analgesics, change from low to high pressure".

baseline 'best'

AIC = ?

cut pts.

	baseline est.	baseline s.e.	best est.	best s.e.
α_2	0.588	0.046	0.588	0.046
α_3	1.136	0.069	1.136	0.069
α_4	1.786	0.079	1.787	0.080
α_5	2.505	0.109	2.506	0.111
intercept	−0.474	0.226	−0.522	0.223
university	−0.523	0.172	−0.523	0.174
analgesics	0.558	0.202	0.561	0.205
change2	—	—	0.031	0.051
change3	—	—	0.164	0.053
γ_F	0.415	0.094	0.424	0.094
γ_T	0.556	0.030	0.557	0.030
$\gamma_T - \gamma_F$	0.142	0.098	0.133	0.098
σ^2	0.566	0.110	0.564	0.111

n.s.

Example: Spatial extremes

Davison et al 2012

- vector observations $(X_{1i}, \ldots, X_{mi}),\ i = 1, \ldots, n$
- example rainfall at each of m locations
- component-wise maxima $Z_1, \ldots, Z_m; Z_j = \max(X_{j1}, \ldots, X_{jn})$
- Z_j are transformed (centered and scaled)

Example: Spatial extremes

- vector observations $(X_{1i}, \ldots, X_{mi}),\ i = 1, \ldots, n$
- example rainfall at each of m locations
- component-wise maxima $Z_1, \ldots, Z_m; Z_j = \max(X_{j1}, \ldots, X_{jn})$
- Z_j are transformed (centered and scaled)

- general theory says

$$\Pr(Z_1 \leq z_1, \ldots, Z_m \leq z_m) = \underline{\exp\{-V(z_1, \ldots, z_m)\}}$$

- function $V(\cdot)$ can be parameterized via Gaussian process models

Example: Spatial extremes

- vector observations $(X_{1i}, \ldots, X_{mi}),\ i = 1, \ldots, n$
- example rainfall at each of m locations
- component-wise maxima $Z_1, \ldots, Z_m; Z_j = \max(X_{j1}, \ldots, X_{jn})$
- Z_j are transformed (centered and scaled)

- general theory says

$$\Pr(Z_1 \le z_1, \ldots, Z_m \le z_m) = \exp\{-V(z_1, \ldots, z_m)\}$$

$h^T \Omega h$

corr = matrix

- function $V(\cdot)$ can be parameterized via Gaussian process models

- example

$$V(z_1, z_2) = z_1^{-1}\Phi\{(1/2)a(h) + a^{-1}(h)\log(z_2/z_1)\} + z_2^{-1}\Phi\{(1/2)a(h) + a^{-1}(h)\log(z_1/z_2)\}$$

$Z(h) = (z_1, z_2), Z(0) = (0, 0), a(h) = h^T\Omega^{-1}h$

$$\Pr(Z_1 \le z_1, \ldots, Z_d \le z_m) = \exp\{-V(z_1, \ldots, z_m)\}$$

- to compute log-likelihood function, need the density
- combinatorial explosion in computing joint derivatives of $V(\cdot)$

D = 10, one likelihood eval is a sum over 100,000 terms

… spatial extremes

$$\Pr(Z_1 \le z_1, \ldots, Z_d \le z_m) = \exp\{-V(z_1, \ldots, z_m)\}$$

- to compute log-likelihood function, need the density
- combinatorial explosion in computing joint derivatives of $V(\cdot)$

D = 10, one likelihood eval is a sum over 100,000 terms

- Davison et al. (2012, *Statistical Science*) used pairwise composite likelihood
- compared the fits of several competing models, using AIC analogue described above
- applied to annual maximum rainfall at several stations near Zurich

Davison et al, 2012

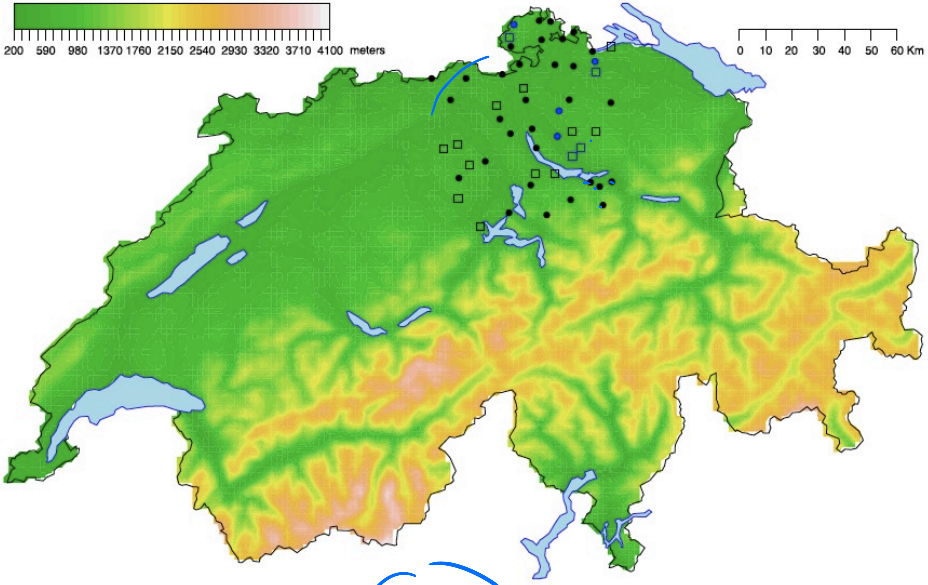


FIG. 1. *Map of Switzerland showing the stations of the 51 rainfall gauges used for the analysis, with an insert showing the altitude. The 36 stations marked by circles were used to fit the models, and those marked with squares were used to validate the models. Data for the pairs of stations with blue symbols appear in Figure 2.*

… Davison et al, 2012

MODELING OF SPATIAL EXTREMES 175

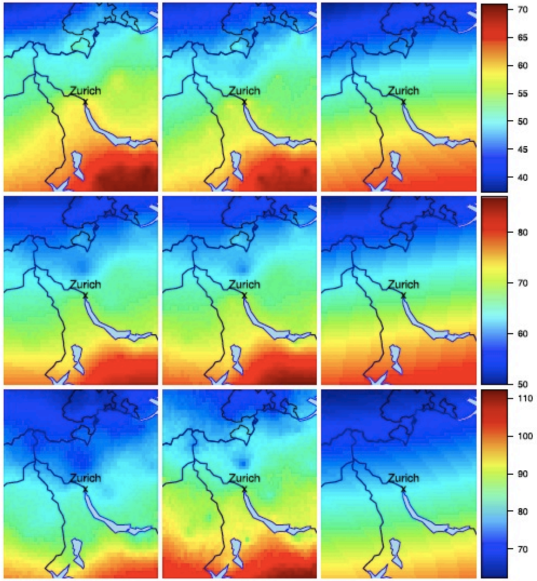


FIG. 3. *Maps of the (predictive) pointwise 25-year return level estimates for rainfall (mm) obtained from the latent variable and max-stable models. The top and bottom rows show the lower and upper bounds of the 95% pointwise credible/confidence intervals. The middle row shows the predictive pointwise posterior mean and pointwise estimates. The left column corresponds to the latent variable model assuming* Gamma(5, 3) *prior on* λ. *The middle column assumes the less informative priors* $\lambda_\eta \sim \mathrm{Gamma}(1, 100)$, $\lambda_\tau \sim \mathrm{Gamma}(1, 10)$ *and* $\lambda_\xi \sim \mathrm{Gamma}(1, 10)$. *The right column corresponds to the extremal t copula model.*

Example: Ising model

Ising model:

$$f(y;\theta) = \exp\Big(\sum_{(j,k)\in E} \theta_{jk} y_j y_k\Big)\frac{1}{Z(\theta)} \qquad j,k = 1,\dots,K$$

Xue et al., 2012

Ravikumar et al., 2010

Example: Ising model

Ising model:

$$f(y;\theta) = \exp\Big(\sum_{(j,k)\in E} \theta_{jk} y_j y_k\Big)\frac{1}{Z(\theta)} \qquad j,k = 1,\dots,K$$

neighbourhood contributions

$$f(y_j \mid y_{(-j)};\theta) = \frac{\exp(2y_j \sum_{k\neq j}\theta_{jk}y_k)}{\exp(2y_j \sum_{k\neq j}\theta_{jk}y_k) + 1} \quad = \quad \exp \ell_j(\theta; y)$$

Xue et al., 2012

Ravikumar et al., 2010

Example: Ising model

Ising model:

$$f(y;\theta) = \exp\Big(\sum_{(j,k)\in E} \theta_{jk} y_j y_k\Big)\frac{1}{Z(\theta)} \qquad j,k = 1,\ldots,K$$

neighbourhood contributions

$$f(y_j \mid y_{(-j)};\theta) = \frac{\exp(2y_j \sum_{k\neq j}\theta_{jk}y_k)}{\exp(2y_j \sum_{k\neq j}\theta_{jk}y_k) + 1} = \exp \ell_j(\theta; y)$$

penalized CL estimation based on sample $y^{(1)},\ldots,y^{(n)}$

$$\max_\theta \left\{ \sum_{i=1}^n \sum_{j=1}^K \ell_j(\theta; y^{(i)}) - \sum_{j<k} P_\lambda(|\theta_{jk}|) \right\}$$

Xue et al., 2012

Ravikumar et al., 2010

Some surprises

- Godambe information $G(\theta)$ can decrease as more component CLs are added

Some surprises

- Godambe information $G(\theta)$ can decrease as more component CLs are added
- pairwise CL can be less efficient than independence CL
- this can't always be fixed by weighting

Xu, 12

Some surprises

- Godambe information $G(\theta)$ can decrease as more component CLs are added
- pairwise CL can be less efficient than independence CL
- this can't always be fixed by weighting

Xu, 12

- parameter constraints can be important
  - Example: binary vector Y,

    $$P(Y_j = y_j, Y_k = y_k) \propto \frac{\exp(\beta y_j + \beta y_k + \theta_{jk} y_j y_k)}{\{1 + \exp(\beta y_j + \beta y_k + \theta_{jk} y_j y_k)\}}$$

  - this model is inconsistent

need $\theta_{jk} \equiv \theta$

Some surprises

#J'H

- Godambe information $G(\theta)$ can decrease as more component CLs are added
- pairwise CL can be less efficient than independence CL
- this can't always be fixed by weighting

Xu, 12

- parameter constraints can be important
  - Example: binary vector Y,

$\beta_1 \quad \beta_2$ $\quad (\beta, \underline{\theta})$

$$P(Y_j = y_j, Y_k = y_k) \propto \frac{\exp(\beta y_j + \beta y_k + \theta_{jk} y_j y_k)}{\{1 + \exp(\beta y_j + \beta y_k + \theta_{jk} y_j y_k)\}}$$

  - this model is inconsistent

need $\theta_{jk} \equiv \theta$

- parameters may not be identifiable in the CL, even if they are in the full likelihood

Yi, 12

$f_2(y_j, y_k) \qquad f_2(y_j \mid y_k)$

- Hammersley-Clifford theorem for conditionals; nothing similar (?) for marginals

when does a set of conditional densities determine a valid joint density

Quasi-likelihood

McCullagh 1983

canonical

- Recall: generalized linear model $y_1, \dots, y_n$ independent, with

$$f(y_i \mid x_i; \beta, \phi) = \exp[\{y_i\theta_i - c(\theta_i)\}/\phi + h(y_i, \phi)]$$

scale

cumulant fn

- ϕ a scale parameter in this exponential family
- $\mathrm{E}(y_i) = \mu_i = c'(\theta_i)$
- $\mathrm{var}(y_i) = \phi V(\mu_i) = \phi c''(\theta_i)$

variance fn

variance function

Quasi-likelihood

- Recall: generalized linear model $y_1, \ldots, y_n$ independent, with

$$f(y_i \mid x_i; \beta, \phi) = \exp[\{y_i\theta_i - c(\theta_i)\}/\phi + h(y_i, \phi)]$$

- ϕ a scale parameter in this exponential family
- $\mathrm{E}(y_i) = \mu_i = c'(\theta_i)$
- $\mathrm{var}(y_i) = \phi V(\mu_i) = \phi c''(\theta_i)$ variance function
- $\boxed{g(\mu_i) = x_i^{\mathrm{T}}\beta}$ link function
- link function converts $\theta_{n\times 1}$ to $\beta_{p\times 1}$
- Standard $V(\mu)$: Normal– 1; Gamma– μ^2; Poisson– μ; Bernoulli– $\mu(1-\mu)$

$$\ell(\beta, \phi; y) = \sum_{i=1}^{n}\left\{\frac{y_i\theta_i - c(\theta_i)}{\phi} + \cancel{h(y_i, \phi)}\right\}$$

$N \quad \phi = \sigma^2$

$V(\mu) \equiv 1$

$\theta_1, \ldots \theta_n$

$\mu_i = c'(\theta_i)$

$y \sim Po(\mu)$

$var(y) = \mu$

$\phi = 1$

$V(\mu) = \mu$

$g(c'(\theta_i)) = x_i^T\beta$

… quasi-likelihood

- log-likelihood

$$\ell(\beta, \phi; y) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - c(\theta_i)}{\phi} + h(y_i, \phi) \right\}$$

- score function

$$\frac{\partial \ell}{\partial \beta_r} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_r} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_r}$$

$\partial / \partial \beta$

$\frac{y_i - \mu_i}{V(\mu_i)} \cdot \{ \downarrow \}$

- MLE

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{x_{ir}}{g'(\mu_i)} = 0$$

2nd

- Bartlett identity:

$$E_{\beta}\left\{ \frac{\partial \ell^2}{\partial \beta_r \partial \beta_s} + \frac{\partial \ell}{\partial \beta_r} \frac{\partial \ell}{\partial \beta_s} \right\} = 0$$

- Suppose instead of a generalized linear model, we had only a partially specified model:

$$E(y_i) = \mu_i, \ \mathrm{var}(y_i) = \phi V(\mu_i), \ g(\mu_i) = x_i^{\mathrm{T}}\beta$$

μ^4

- $g(\cdot), V(\cdot)$ known
- unbiased estimating equation

$$g(y;\beta) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{x_{ir}}{g'(\mu_i)}$$

$= 0$: define $\tilde\beta$

$\psi(y;\beta)$

not $g(\mu)$

- if $g(y;\tilde\beta) = 0$, then asymptotic variance of $\tilde\beta$ is

$$\underbrace{E\left\{-\frac{\partial g(y;\beta)}{\partial \beta^{\mathrm{T}}}\right\}^{-1}}_{H^{-1}} \underbrace{\mathrm{var}\{g(y;\beta)\}}_{J} \underbrace{E\left\{-\frac{\partial g(y;\beta)}{\partial \beta}\right\}^{-1}}_{H^{-1}}$$

as with composite likelihood

for any $\psi(y;\beta)$

$E_{\beta}\{\psi(y;\beta)\} = 0$

$\tilde\beta := \psi(y,\tilde\beta) = 0$

... quasi-likelihood

- With $g(y;\beta) = \Sigma g_i(y_i;\beta) = \sum \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{x_{ir}}{g'(\mu_i)}$

$\psi(y;\beta)$

$E(Y_i)$, $\mathrm{var}(Y_i)$

- $\boxed{E\left\{-\frac{\partial g(y;\beta)}{\partial \beta^{\mathrm{T}}}\right\}} = \sum_{i=1}^{n} x_i x_i^{\mathrm{T}} \frac{1}{g'(\mu_i)^2 \phi V(\mu_i)} = \phi^{-1} X^{\mathrm{T}} W X = \mathrm{var}\{g(y;\beta)\}$

special est'g eq: where $H = J$

- $W = \mathrm{diag}(w_j)$, $\quad w_j = \{g'(\mu_j)^2 V(\mu_j)\}^{-1}$, $j = 1, \dots, n$

- quasi-likelihood function

$$Q(\beta; y) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - u}{\phi V(u)} du$$

quasi-lik.

$Q: \mathbb{R}^d \to \mathbb{R}^1$

$Q'(\beta) = 0 \longrightarrow$ usual GLM type

- this only works for models of this form
- called quasi-likelihood because $\partial Q/\partial\beta$ gives estimating equation with expected value 0, and 2nd Bartlett identity holds

log-

$E\{Q''(\beta) + \mathrm{var}\, Q'(\beta)\} = 0$

Longitudinal data

- suppose now our observations come in groups: $y_{ij}, j = 1, \ldots, m_i; i = 1, \ldots, n$
- could be repeated measurements on subjects
- or measurements of members of the same cluster/family/group

- assume GLM-type structure $E(y_{ij}) = \mu_{ij}$, $g(\mu_{ij}) = x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}} b_i$
- random effects b_i induce correlation among observations in the same group; e.g. assume $b_i \sim N(0, \Omega_b)$

random effects

corr $\doteq$

- GLM variance structure $\mathrm{var}(y_{ij}) = V_i(\beta, \alpha)$ for example
- α are extra parameters in the variance-covariance matrix
- QL-type estimating equations

$g(\mu) = x^T\beta$

$\mathrm{var}\, y = V(\mu) = V(\beta)$

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta^{\mathrm{T}}}\right)^{\mathrm{T}} V_i^{-1}(\alpha, \beta)(y_i - \mu_i) = 0$$

$g'(\mu_i)\, x_i$

y: μ:

Generalized Estimating Equations

Liang & Zeger, 1986

GEE

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta^{\mathrm{T}}} \right)^{\mathrm{T}} V_i^{-1}(\alpha, \beta)(y_i - \mu_i) = 0$$

- parameter α in variance function doesn't divide out, as in univariate case
- we will need an estimate $\hat{\alpha}$ from somewhere

many suggestions in the literature

- Liang & Zeger suggested using a "working covariance matrix" to get an estimate of β
- e.g. could assume independence, or $AR(1)$, or ...
- estimates of β will still be consistent, but the asymptotic variance will be of the sandwich form as the model is misspecified
- there is no integrated function that serves as a quasi-likelihood in this setting