Various 'types' of likelihood

1. likelihood, marginal likelihood, conditional likelihood, profile likelihood, adjusted profile likelihood

2. semi-parametric likelihood, partial likelihood

3. empirical likelihood, penalized likelihood

4. quasi-likelihood, composite likelihood

5. simulated likelihood, indirect inference

6. bootstrap likelihood, *h*-likelihood, weighted likelihood, pseudo-likelihood, local likelihood, sieve likelihood

November 13

- comment on report and presentation
- mis-specified models
- composite likelihood
- quasi-likelihood
- simulation likelihood

Exercises November 6 **STA 4508S (Fall, 2018)**

1. Re-submit the 2nd set of exercises (if you wish) in light of the corrections and notes presented in class on November 6.

2. Choose a paper for presentation on November 27, and provide the complete citation and a one-sentence description of the paper.

   You should plan for a 15 minute presentation followed by 5 minutes of questions. The presentation can be either on the blackboard or on slides. My guideline for number of slides is one per minute. But be warned you need to talk a lot to have one standard-looking slide take one minute to present.

   Your report would I expect be between five and ten pages (five is enough), and would include:

   (a) Complete citation

   (b) One paragraph overview of the main problem and results (without quoting the abstract).

   (c) A paragraph setting the work in context of the literature and the authors' previous work. (Again, not quoted directly from the paper.)

   (d) One or two sections outlining the techniques used to get from problem to results. For some papers this might mean quoting the most important theorem(s) and summarizing the key idea of the proof(s); for others it might be describing a new model that is proposed and explaining the key new features of this.

   (e) Two or three paragraphs providing your assessment of the paper and the work: was it interesting? was the paper well written? does it suggest further work (different from what is outlined in the 'further work' section by the authors)? Is there an application suggested, and if so, did you find it convincing? How does it tie into topics discussed in this course?

- $y_1, \ldots, y_n$ 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 $G(\cdot)$ in the family $\{f(\cdot;\theta), \theta \in \Theta\}$

... misspecified models

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

- $\theta^* = \arg\min_\theta KL(\theta)$
- $\theta^* = \arg\max_\theta \int \log\{f(y;\theta)\}g(y)dy = \arg\max_\theta \mathrm{E}_G\{\ell(\theta;y)\}$
- leads to a proof that the maximum likelihood estimator converges to θ^* under some smoothness conditions, etc.
- if $g(y) = f(y;\theta_0)$, then $\theta^* = \theta_0$ true density is in the model family
- otherwise θ^* is the 'least false' parameter value

... misspecified models

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

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

... 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, 1982

- wthis 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: $(\partial/\partial\theta)\ell(\hat{\theta}; y) = 0$
- 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^*
- $E_G(-\partial^2/\partial\theta^2)\ell(\theta; y) = \int (-\partial^2/\partial\theta^2)\ell(\theta; y)g(y)dy \equiv H(\theta)$ $= H_G(\theta)$
- $E_G\{(\partial/\partial\theta)\ell(\theta; y)\}^2 = \int \{(\partial/\partial\theta)\ell(\theta; y)\}^2 g(y)dy \equiv J(\theta) \neq H(\theta)$
- $(\hat{\theta} - \theta^*) = H(\theta^*)^{-1}U(\theta^*)\{1 + o_p(1)\}$ $U(\theta) = (\partial/\partial\theta)\ell(\theta)$
- $\hat{\theta} \dot{\sim} N\{\theta^*, \mathcal{G}^{-1}(\theta^*)\}$ $\mathcal{G}(\theta) = H(\theta)J^{-1}(\theta)H(\theta)$

Examples

- $y_1, \ldots, y_n$ i.i.d. $\sim G$; we assume $f(y;\theta)$ is $N(\mu, \sigma^2)$
- $\hat{\mu} = \bar{y}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$ $\mu^* = \mu_G, \sigma^{*2} = \sigma_G^2$
- $\partial_\theta \ell(\theta; y) = [\Sigma(y_i - \mu)/\sigma^2, \quad -(n/2\sigma^2) + \Sigma(y_i - \mu)^2/(2\sigma^4)]$
- $H(\theta) = n \begin{bmatrix} 1/(\sigma_G^2) & 0 \\ 0 & 1/(2\sigma_G^4) \end{bmatrix}$ $H(\theta) = -\mathrm{E}_G \ell''(\theta; y)$

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

  $J(\theta) = n \begin{bmatrix} 1/(\sigma_G^2) & \mu_3/(2\sigma_G^6) \\ \mu_3/(2\sigma_G^6) & (\mu_4 - \sigma_G^4)/(4\sigma_G^8) \end{bmatrix}$ ntbc
- $\sigma_G^2 = \mathrm{var}_G(y_1)$, $\mu_3 = \mathrm{E}_G(y_1 - \mu_G)^3$, $\mu_4 = \mathrm{E}_G(y_1 - \mu_G)^4$
- $\mathcal{G}(\theta_G) = n^{-1} \begin{bmatrix} \sigma_G^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma_G^4 \end{bmatrix}$ not uncorrelated

  ntbc

... examples

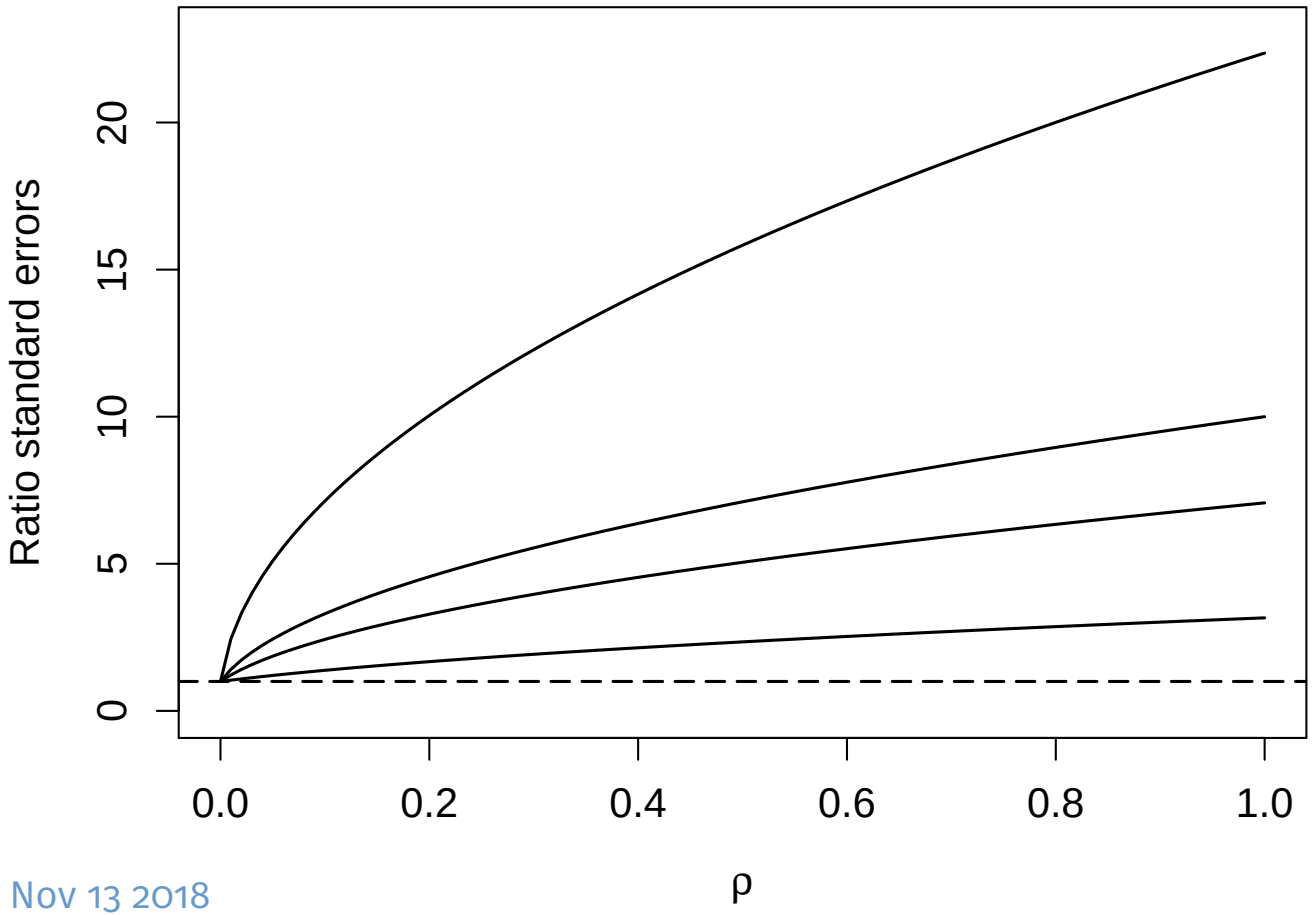
- linear regression model G: $y = X\beta_0 + \epsilon, \quad \epsilon \sim (0, \sigma^2 R)$

  linear regression; correlated errors
- working model $\quad f(y;\beta) = N(X\beta, \sigma^2 I)$ uncorrelated errors
- $\hat{\beta} = (X^T X)^{-1} X^T y$ least squares estimator; mle under normality
- $E_G(\hat{\beta}) = \beta_0$ LSE unbiased (consistent)
- $\text{var}_G(\hat{\beta}) = \sigma^2 (X^T X)^{-1} X^T R X (X^T X)^{-1}$ sandwich variance
- working model variance is incorrect $\sigma^2 (X^T X)^{-1}$
- single intercept model $\hat{\beta} = \bar{y}$, $R_{ij} = \rho$, $\text{var}(\bar{y}) = (\sigma^2/n)\{1 + \rho(n-1)\}$

  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 3.3



Composite likelihood: recap

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

- Composite Log-Likelihood: Lindsay, 1988

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

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

- also called:
  - pseudo-likelihood (spatial modelling)
  - quasi-likelihood (econometrics)
  - limited information method (psychometrics)

... composite likelihood: recap

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

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	$G(\theta) = G_n(\theta) = H_n(\theta) J_n^{-1}(\theta) H_n(\theta) = O(n)$

... composite likelihood: recap

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

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

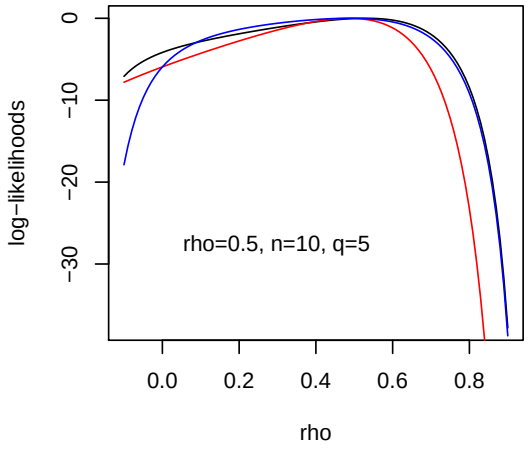
... inference

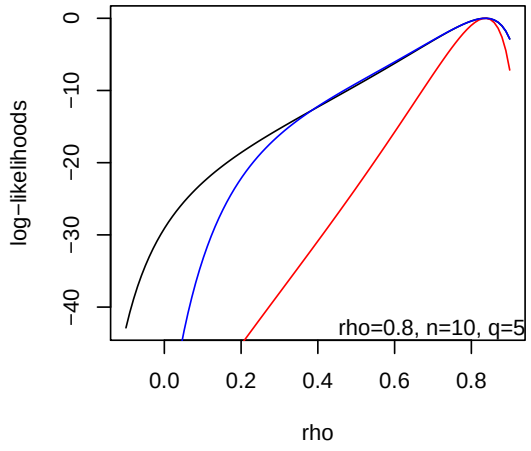
- $\tilde{w}_e(\theta) = (\hat{\theta}_{CL} - \theta)^{\mathrm{T}} G(\theta)(\hat{\theta}_{CL} - \theta) \xrightarrow{d} \chi^2_p$
- $\tilde{w}_u(\theta) = U(\theta)^{\mathrm{T}} G^{-1}(\theta) U(\theta) \xrightarrow{d} \chi^2_p$
- $\tilde{w}(\theta) = 2\{\mathrm{c}\ell(\hat{\theta}_{CL}) - \mathrm{c}\ell(\theta)\} \doteq (\hat{\theta}_{CL} - \theta)^{\mathrm{T}} i(\theta)(\hat{\theta}_{CL} - \theta)$

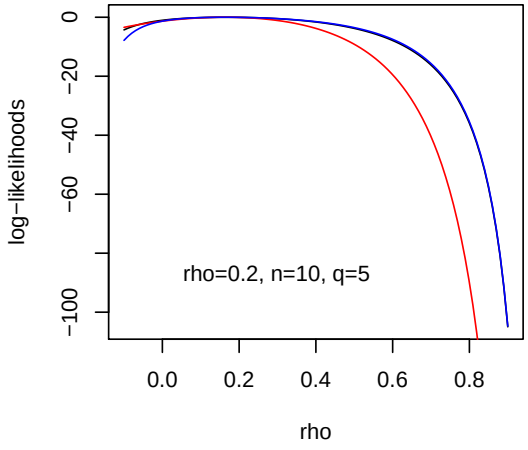
$$\xrightarrow{d} \sum_{a=1}^{d} \mu_a Z_a^2 \quad Z_a \sim N(0,1)$$

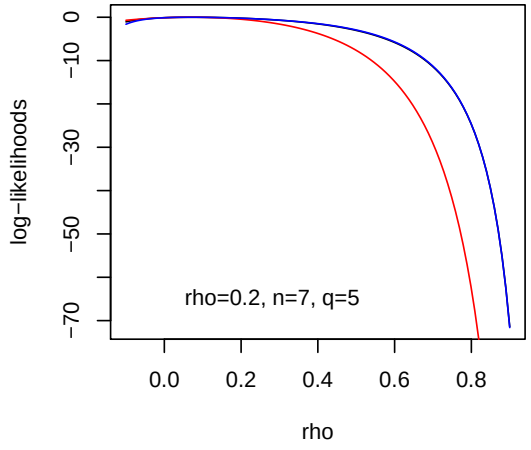
- $\mu_1, \ldots, \mu_d$ are the eigenvalues of $J(\theta)H(\theta)^{-1}$ — non-central χ^2 limit
- $J(\theta) = \mathrm{var} U(\theta), \qquad H(\theta) = -\mathrm{E}\partial U(\theta)/\partial\theta^{\mathrm{T}}$
- if $J(\theta) = H(\theta)$, $w(\theta) \dot{\sim} \chi^2_d$
- if $d = 1$, $w(\theta) \dot{\sim} \mu_1 \chi^2_1 = J(\theta)H^{-1}(\theta)\chi^2_1$ — H, J both scalars

Likelihood ratio test









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
- $w(\psi) = 2\{c\ell(\hat{\theta}_{CL}) - c\ell(\tilde{\theta}_\psi)\} \dot{\sim} \sum_{a=1}^{d_0} \mu_a Z_a^2$
- $\mu_1, \ldots, \mu_{d_0}$ are the eigenvalues of $(H^{\psi\psi})^{-1}G^{\psi\psi}$

Kent, 1982

- Godambe information needs to be estimated

$$\begin{aligned} \hat{H}(\theta) &= -\partial^2 c\ell(\hat{\theta}_{CL})/\partial\theta\partial\theta^{\mathrm{T}} \\ \hat{J}(\theta) &= n^{-1}\sum_{i=1}^{n} U_{CL}(\theta; y_i)U_{CL}^{\mathrm{T}}(\theta; y_i) \end{aligned}$$

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

Model selection

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

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

- 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

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

used for selection of tuning parameters

Example: CL with dichotomized MV Normal

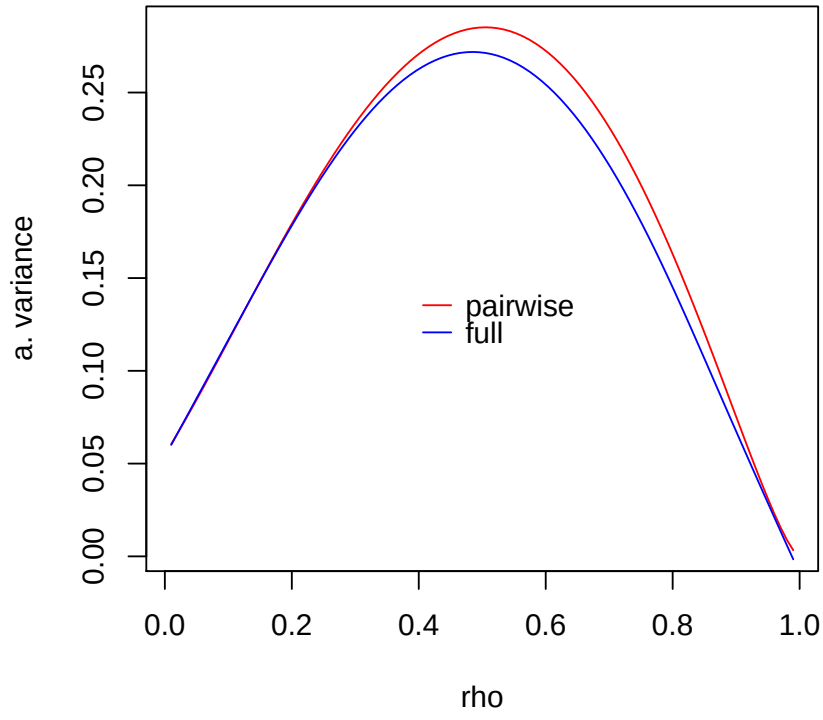
$$Y_{ir} = 1\{Z_{ir} > 0\} \qquad Z \sim N(0, R) \qquad r = 1, \ldots, m; i = 1, \ldots, n$$

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

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

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

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



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

Numbers incorrect in Cox & Reid 2004 Table 1

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: $Pr(y_{ir} = 1 \mid b_i) = \Phi(x'_{ir}\beta + b_i); \quad b_i \sim N(0, \sigma_b^2)$

- 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 $\Phi_2(\cdot, \cdot; \rho_{irs})$

... 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

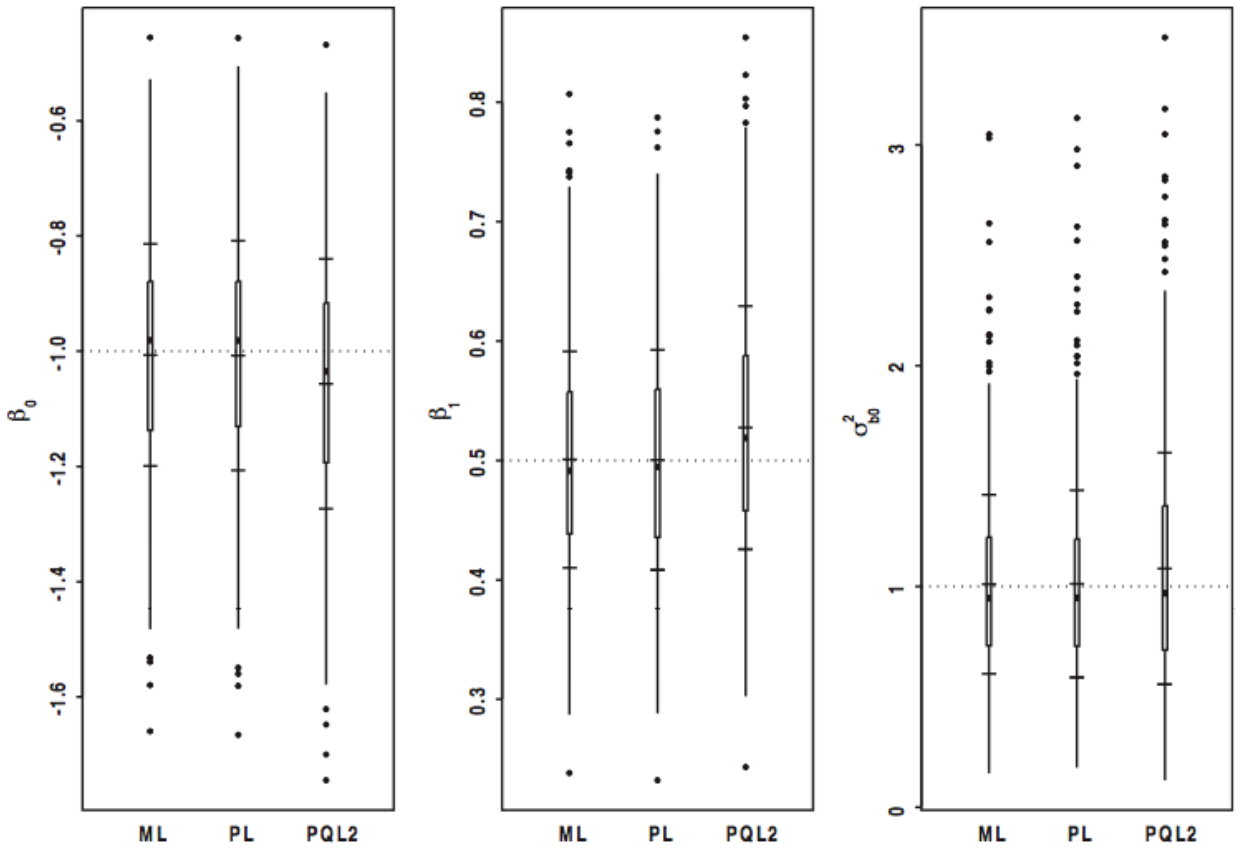


Fig. 5. Boxplots of ML, PL and PQL2 simulated parameter estimates under Model (10) with random intercept.

Example: longitudinal count data

Henderson & Shimura, 2003

- 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 $\rho = 0$

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 $\quad y_{ij}^* = x_{ij}^{\mathrm{T}}\beta + u_i + \epsilon_{ij}$
- $y_{ij} = k \Leftrightarrow a_{k-1} < y_{ij}^* < a_k$
- if $u_i \sim N(0, \sigma^2)$ and $\epsilon_{ij} \sim N(0, 1)$

$$L(\theta; y) = \prod_{i=1}^{n} f(y_{i1}, \ldots, y_{im_i}) =$$
$$\prod_{i=1}^{n} \int_{-\infty}^{\infty} \prod_{j=1}^{m_i} \{\Phi(a_{y_{ij}} - x_{ij}^{\mathrm{T}}\beta - u_i) - \Phi(a_{y_{ij}-1} - x_{ij}^{\mathrm{T}}\beta - u_i)\}\phi(\tfrac{u_i}{\sigma})du_i$$

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

... 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 $\text{corr}(\epsilon_{ij}, \epsilon_{ij'}) = \exp(\delta|t_{ij} - t_{ij'}|)$
- now

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

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

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

- pairwise log-likelihood:

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

weights are 1 or 0, depending on distance between time 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".

	est.	s.e.	est.	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
$\gamma_{\mathrm{T}} - \gamma_{\mathrm{F}}$	0.142	0.098	0.133	0.098
σ^2	0.566	0.110	0.564	0.111

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)
- general theory says

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

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

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

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

... spatial extremes

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

- 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

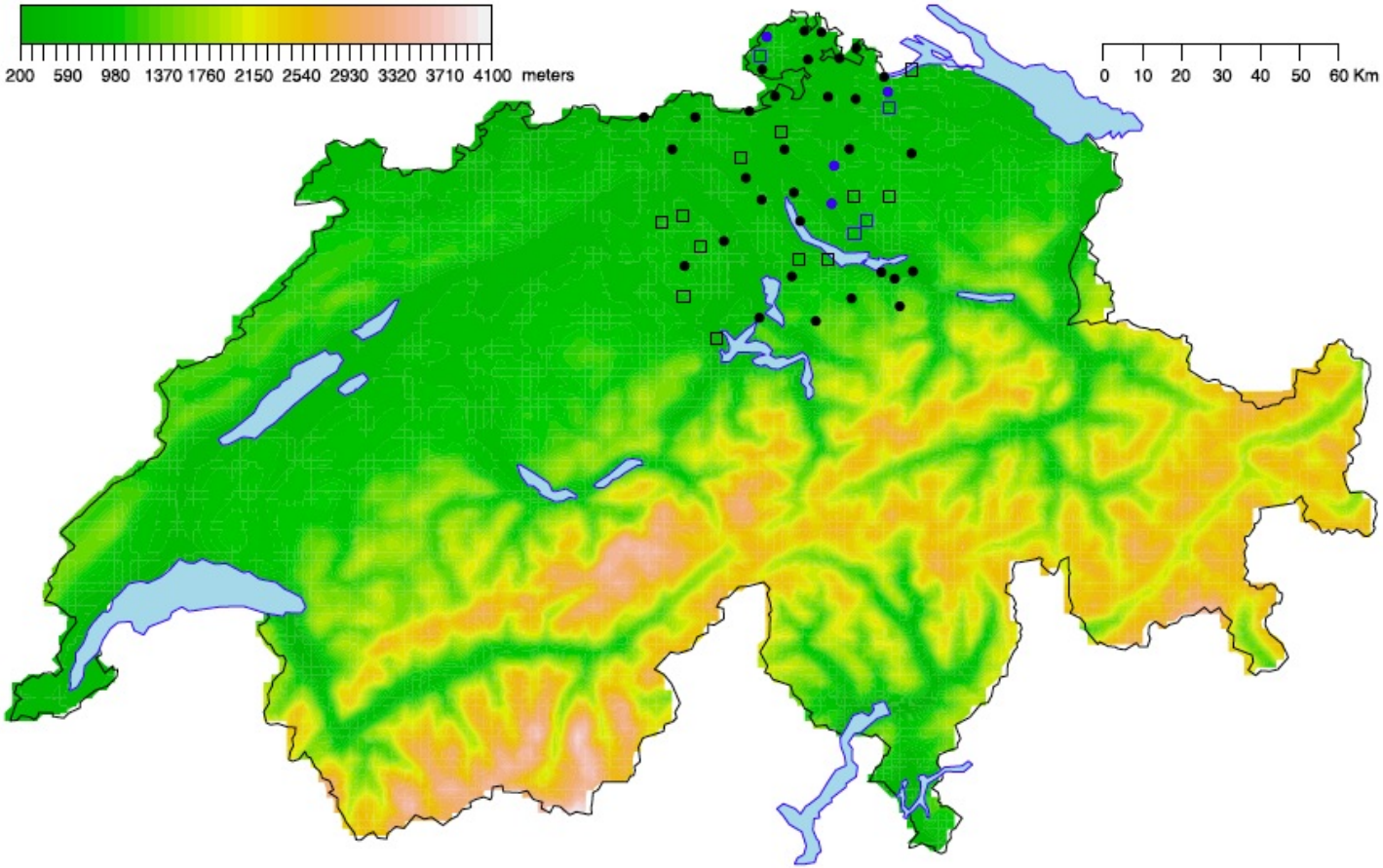


FIG. 1. *Map of Switzerland showing the stations of the 51 rainfall gauges used for the analysis, with an insert showing the altitude. The 36 stations marked by circles were used to fit the models, and those marked with squares were used to validate the models. Data for the pairs of stations with blue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

… Davison et al, 2012

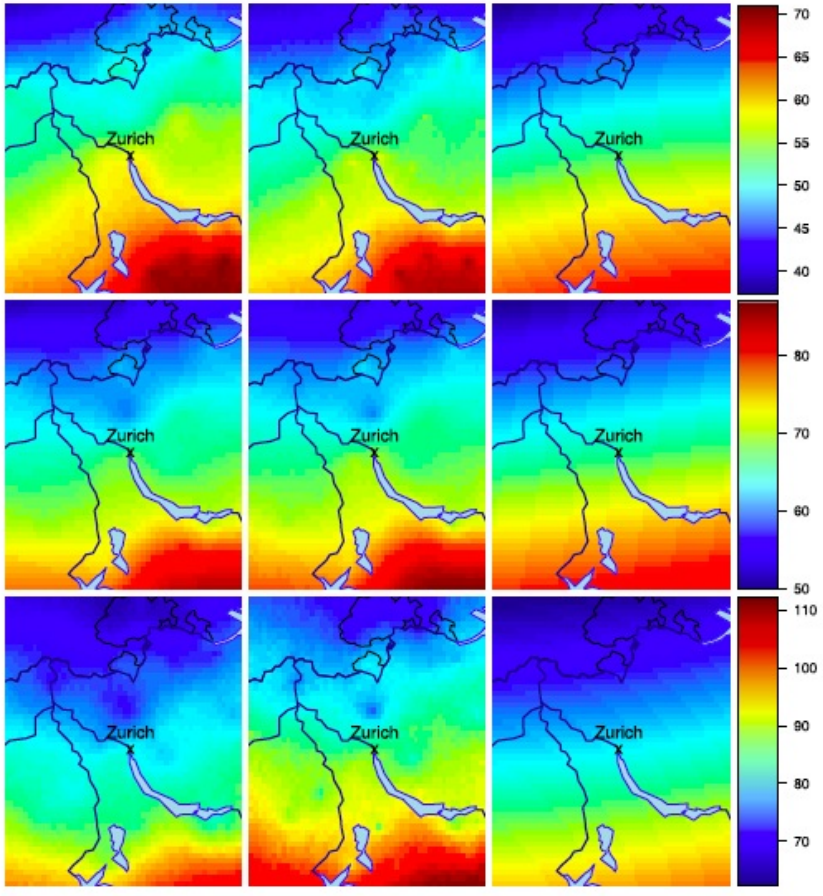


FIG. 3. *Maps of the (predictive) pointwise 25-year return level estimates for rainfall (mm) obtained from the latent variable and max-stable models. The top and bottom rows show the lower and upper bounds of the 95% pointwise credible/confidence intervals. The middle row shows the predictive pointwise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

Example: Ising model

Ising model:

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

neighbourhood contributions

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

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

Composite likelihood: recap

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

- Composite Log-Likelihood: Lindsay, 1988

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

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

- choice of weights and choice of sets $\{\mathcal{A}_{||}\}$
- choice of weights generally problem specific

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 $\theta_{jk} \equiv \theta$
- parameters may not be identifiable in the CL, even if they are in the full likelihood Yi, 12

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

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