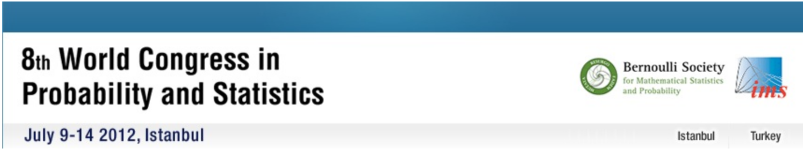
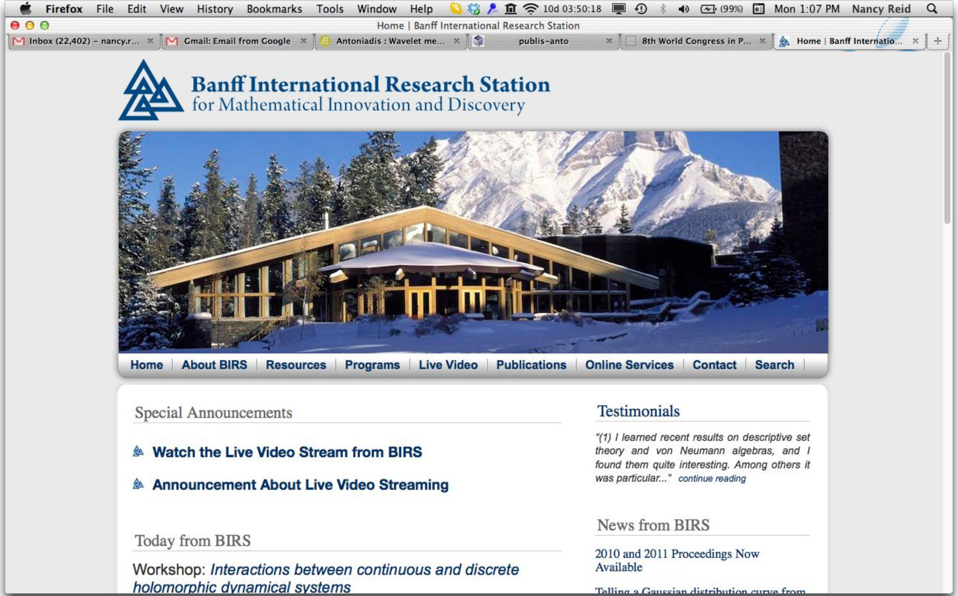
Composite Likelihood

Nancy Reid

July 12, 2012

with Harry Joe, Cristiano Varin
and thanks to Don Fraser, Grace Yi, Ximing Xu




Firefox File Edit View History Bookmarks Tools Window Help
10d 03:50:18
(99%)
Mon 1:07 PM Nancy Reid
Home | Banff International Research Station
Inbox (22,402) - nancy.r...
Gmail: Email from Google
Antoniadis : Wavelet me...
publis-anto
8th World Congress in P...
Home | Banff Internatio...
Banff International Research Station
for Mathematical Innovation and Discovery
Home
About BIRS
Resources
Programs
Live Video
Publications
Online Services
Contact
Search
Special Announcements
Watch the Live Video Stream from BIRS
Announcement About Live Video Streaming
Today from BIRS
Workshop: Interactions between continuous and discrete holomorphic dynamical systems
Testimonials
"(1) I learned recent results on descriptive set theory and von Neumann algebras, and I found them quite interesting. Among others it was particular..." continue reading
News from BIRS
2010 and 2011 Proceedings Now Available

Terminology

- Model $Y \sim f(y;\theta), \quad y \in \mathbb{R}^m, \quad \theta \in \mathbb{R}^p$
- Events $A_1, \ldots, A_k$; "sub-densities" $f(y \in A_k;\theta)$
- Composite log-likelihood

$$c\ell(\theta; y) = \sum_{k=1}^{K} w_k \log f(y \in A_k;\theta) = \sum_{i=1}^{K} w_k\, \ell(\theta; y \in A_k)$$

- w_k weights to be determined
- composite likelihood is a type of:
  - pseudo-likelihood (spatial modelling);
  - quasi-likelihood (econometrics);
  - limited information method (psychometrics)
  - ...

Examples of $c\ell(\theta)$

$$\sum_{r=1}^{m} w_r \log f_1(y_r;\theta) \quad \text{Independence}$$

$$\sum_{r=1}^{m} \sum_{s>r} w_{rs} \log f_2(y_r, y_s;\theta) \quad \text{Pairwise}$$

$$\sum_{r=1}^{m} w_r \log f(y_r \mid y_{(-r)};\theta) \quad \text{Conditional}$$

$$\sum_{r=1}^{m} \sum_{s>r} w_{rs} \log f(y_r \mid y_s;\theta) \quad \text{All pairs conditional}$$

$$\sum_{r=1}^{m} w_r \log f(y_r \mid y_{r-1};\theta) \quad \text{Time series}$$

$$\sum_{r=1}^{m} w_r \log f(y_r \mid \text{`neighbours' of } y_r;\theta) \quad \text{Spatial}$$

likelihood of (small) blocks of observations; pretend blocks indep.

likelihood of pairwise differences

your favourite fix here ...

Inference

- Sample $y_1, \dots, y_n$ independent
- Composite log-likelihood $\sum_{i=1}^{n} c\ell(\theta; y_i)$; maximized at $\hat{\theta}_{CL}$
- As $n \longrightarrow \infty$:

$$\sqrt{n}(\hat{\theta}_{CL} - \theta) \quad \xrightarrow{\mathcal{L}} \quad N\{0, G^{-1}(\theta)\},$$

- Godambe information $G(\theta) = H(\theta)J^{-1}(\theta)H(\theta)$
- $H(\theta) = \mathrm{E}\left\{-\dfrac{\partial^2 c\ell(\theta; Y_i)}{\partial\theta\partial\theta^T}\right\}, \quad J(\theta) = \mathrm{var}\left\{\dfrac{\partial c\ell(\theta; Y_i)}{\partial\theta}\right\}$

... inference

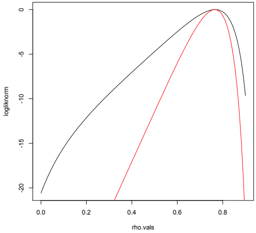
- Sample $y_1, \ldots, y_n$ independent
- Composite log-likelihood $c\ell^n(\theta) = \sum_{i=1}^{n} c\ell(\theta; y_i)$;
- CL log-likelihood ratio $w_{CL}(\theta) = 2\{c\ell^n(\hat{\theta}_{CL}) - c\ell^n(\theta)\}$
- As $n \longrightarrow \infty$:

$$w_{CL}(\theta) \quad \xrightarrow{\mathcal{L}} \quad \sum_{j=1}^{p} \lambda_j \chi^2_{1j}$$

- λ_j eigenvalues of $J^{-1}(\theta)H(\theta)$

What do we know?

- $\hat{\theta}_{CL}$ not fully efficient, unless $G(\theta) = H(\theta)J^{-1}(\theta)H(\theta) = i(\theta)$
- $c\ell(\theta)$ is not a log-likelihood function



- efficiency of $\hat{\theta}_{CL}$ can be pretty high, in many applications
- $w_{CL}(\theta)$ can be re-scaled to $\dot{\sim}\ \chi^2_p$

  Chandler & Bate 07, Salvan et al. 11
- a little about asymptotics as $m \to \infty$, n fixed or increasing slowly

... what do we know?

- careful choice of weights can improve efficiency of $\hat{\theta}_{CL}$
  in special cases

- weights can be used to incorporate sampling information, including missing data

  Yi 12, Molenberghs 12, Briollais & Choi 12

- composite likelihood can be used for model selection

$$
\begin{aligned}
AIC_{CL} &= -2c\ell^n(\hat{\theta}_{CL}) + 2\,\mathrm{tr}\{J(\hat{\theta})H^{-1}(\hat{\theta})\} \\
BIC_{CL} &= -2c\ell^n(\hat{\theta}_{CL}) + \log(n)\,\mathrm{tr}\{J(\hat{\theta})H^{-1}(\hat{\theta})\}
\end{aligned}
$$

- and prediction

- combination of full likelihood for mean parameters and CL for covariance parameters works well in some settings

What don't we know?

- Design
  - marginal vs. conditional
  - choice of weights
  - down-weighting 'distant' observations
  - choosing blocks and block sizes
- Uncertainty estimation
  - $\hat{J}(\hat{\theta}_{CL}) = \widehat{\text{var}}\{\partial c\ell(\theta)/\partial\theta\}$
    need replication; need lots of replication
  - perhaps estimate $G(\hat{\theta}_{CL})$ or $\text{var}(\hat{\theta}_{CL})$ directly – bootstrap, jackknife
  - or estimate using ideas from higher-order asymptotic approximations — Fraser 12
  - or try to find some orthogonal components — Lindsay 12

... what don't we know?

- Identifiability (1): does there exist a model compatible with a set of marginal or conditional densities?
- Identifiability (2): what if different components are estimating different parameters?
- Robustness: CL uses 'low-dimensional' information: is this a type of robustness?
  - find a class of models with same low-d marginals — Xu 12
  - classical perturbation of starting model (using copulas?) — Joe 12
  - random effects models might be amenable to theoretical analysis — Jordan 12
- asymptotic theory for large m (long vectors of responses), small n
- relationship to GEE

Some surprises

- $Y \sim N(\underline{\mu}, \Sigma)$ $\quad - \hat{\mu}_{CL} = \hat{\mu}, \hat{\Sigma}_{CL} = \hat{\Sigma}$ (marginal or conditional (pairwise or full))
- $Y \sim N(\mu \underline{1}, \sigma^2 R), \quad R = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \ddots & \ddots & \vdots \\ \rho & \dots & \rho & 1 \end{pmatrix}$
  - $\hat{\theta}_{CL} = \hat{\theta}, \quad G(\theta) = i(\theta), G(\theta) = H(\theta)J^{-1}(\theta)H(\theta)$
  - $H(\theta) = \text{var(Score)}, J = \text{E}(\nabla_\theta \text{Score}), H \neq J,$
- $Y \sim (0, R)$: $\hat{\rho}_{CL} \neq \hat{\rho}$; $\text{a.var}(\hat{\rho}_{CL}) > \text{a.var}(\hat{\rho})$
- efficiency improvement when nuisance parameter is unknown Mardia et al 08; Xu 12
- CL can be fully efficient, even if $H(\theta) \neq J(\theta)$

... 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
- parameters may not be identifiable in the CL, even if they are in the full likelihood Yi, 12

- CL may help get rid of nuisance parameters (e.g. by conditioning) Hjort and Varin, 07

Some (more) interesting applications

- spatial data and space-time data
  - conditional approaches seem more natural
  - condition on neighbours (in space); some small number of lags (in time)
  - some form of blockwise components often proposed Stein et al, 04; Caragea and Smith, 07
  - fMRI time series Kang et al 12
  - air pollution and health effects Bai et al 12
  - computer experiments: Gaussian process models Xi 12
- spatially correlated extremes
  - joint tail probability known
  - joint density requires combinatorial effort (partial derivatives)
  - composite likelihood based on joint distribution of pairs, triples seems to work well

Davison et al 12; Genton et al 12

... applications

- time series – a case of large m, fixed n
  - need new arguments re consistency, asymptotic normality
  - consecutive pairs: consistent, not asy. normal
  - $AR(1)$: consecutive pairs fully efficient; all pairs terrible (consistent, highly variable)
  - $MA(1)$: consecutive pairs terrible

Davis and Yau 11

- genetics: estimation of recombination rate
  - somewhat similar to time series
  - but correlation may not decrease with increasing length
  - suggesting all possible pairs may be inconsistent
  - joint blocks of short sequences seems preferable
- linkage disequilibrium
- family based sampling

Larribe and Fearnhead 11; Choi and Briollais 12

... applications

- Gaussian graphical models — Gao and Massam 12
  - symmetry constraints have a natural formulation in terms of elements of concentration matrix
  - conditional distribution of $y_j \mid y_{(-j)}$
- multivariate binary data for multi-neuron spike trains — Amari 12
- CL as a working likelihood in 'maximization by parts' — Bellio 12
- latent variable models in psychometrics — Moustaki 12, Maydeu-Olivares 12
- many linear and generalized linear models with random effects
- multivariate survival data
- ...

Some dichotomies

- conditional vs marginal
- pairwise vs everything else
- unstructured vs time series/spatial
- weighted vs unweighted
- "it works" vs "why does it work?" vs "when will it not work"
- ...