# Composite Likelihood

Nancy Reid

January 30, 2012

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

Pattern Theory Group

# Background

- parametric model $f(y;\theta), \quad y \in \mathbb{R}^m; \theta \in \mathbb{R}^d$
- likelihood function $L(\theta; y) \propto f(y;\theta)$
- why likelihood?

- maximum likelihood estimator is consistent and asymptotically efficient
- $\hat{\theta} \dot{\sim} N\{\theta, j^{-1}(\hat{\theta})\}$ $\qquad j(\theta) = -\ell''(\theta); \ell(\theta) = \log L(\theta)$

- likelihood ratio, or log-likelihood difference, captures asymmetry in the model
- $w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \dot{\sim} \chi^2_d$

- combine with prior for Bayesian inference

- EM algorithm for computing maximum likelihood estimate

# ... background

- warning: likelihood methods need regularity conditions on the model
- can have poor finite sample behaviour for large numbers of parameters
- priors also very tricky with large numbers of parameters
- finite sample corrections may be advisable

- difficulty with construction of the likelihood function
  - inversion of large covariance matrices
  - intractable integrals; awkward normalization constants
  - combinatorial explosion
  - nuisance components difficult to specify
- lower dimensional marginal or conditional distributions may be tractable
- combine these to form composite likelihood

# Terminology

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

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

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

# Example: spatial generalized linear models

- generalized linear geostatistical models

$$\mathrm{E}\{Y(s) \mid u(s)\} = g\{x(s)^T\beta + u(s)\}, \quad s \in \mathcal{S} \subset \mathbb{R}^d, d \geq 2$$

Diggle & Ribeiro, 2007

- random intercept $u(\cdot)$ is a realization of a stationary GRF, mean 0, covariance

$$\mathrm{cov}\{u(s), u(s')\} = \sigma^2\rho(s - s'; \alpha)$$

- $m$ observed locations $y = (y_1, \ldots, y_m)$ with $y_i = y(s_i)$
- likelihood function

$$L(\theta; y) = \int_{\mathbb{R}^n} \prod_{i=1}^{m} f(y_i \mid u_i; \theta) \underbrace{f(u; \theta)}_{MVN(0,\Sigma)} du_1 \ldots du_m$$

- no factorization into lower dimensional integrals, as with independent observations from the "‘usual" GLMMs

# ... spatial generalized linear models

- $L(\theta; y) = \int_{\mathbb{R}^n} \prod_{i=1}^{n} f(y_i \mid u_i; \theta) f(u; \theta) du_1 \ldots du_n$
- simulation methods, MCMC, MCEM, etc., costly $O(m^3)$
- pairwise likelihood

$$L_{pair}(\theta; y) = \prod_{\{(i,j) \in \mathcal{S}_\delta\}} \int_{\mathbb{R}^2} f(y_i \mid u_i; \theta) f(y_j \mid u_j; \theta) f_2(u_i, u_j; \theta) du_i du_j$$

Heagerty & Lele (1998), Varin (2008)

- comments:
  - product of bivariate integrals
  - accurate quadrature approximations available
  - use only close pairs: $\mathcal{S}_\delta = \{(i, j) : ||s_i - s_j|| < \delta\}$
  - computational cost $O(n)$

# Composite conditional likelihoods

- Besag (1974) pseudo-likelihood

$$L_C(\theta; y) = \prod_{r=1}^{m} f(y_r \mid \{y_s : y_s \text{ neighbour of } y_r; \theta)$$

- pairwise conditional

$$L_C(\theta; y) = \prod_{r=1}^{m} \prod_{s=1}^{m} f(y_r \mid y_s; \theta)$$

Molenbergs & Verbeke (2005); Mardia et al. (2009)

- full conditional

$$L_C(\theta; y) = \prod_{r=1}^{m} f(y_r \mid y_{(-r)}; \theta)$$

- time series

$$L_C(\theta; y) = \prod_{r=1}^{m} f(y_r \mid y_{r-1}; \theta)$$

# Composite marginal likelihoods

- independence likelihood

$$L_{ind}(\theta; y) = \prod_{r=1}^{m} f(y_r; \theta)$$

Chandler & Bate (2007)

- pairwise likelihood

$$L_{pair}(\theta; y) = \prod_{r=1}^{m} \prod_{s=r+1}^{m} f(y_r, y_s; \theta)$$

Cox & Reid (2004); Varin (2008)

- pairwise differences

$$L_{diff}(\theta; y) = \prod_{r=1}^{m-1} \prod_{s=r+1}^{m} f(y_r - y_s; \theta)$$

Curriero & Lele (1999)

- optimal combination of $L_{ind}$ and $L_{pair}$

# Derived quantities

- composite log-likelihood
  $c\ell(\theta; y) = \log L_C(\theta; y) = \sum_{k=1}^{K} w_k \ell_k(\theta; y)$

- composite score
  $u(\theta; y) = \nabla_\theta c\ell(\theta; y) = \sum_{k=1}^{K} w_k \nabla_\theta \ell_k(\theta; y)$   $E\{u(\theta; Y)\} = 0$

- sensitivity matrix $H(\theta) = \mathrm{E}_\theta\{-\nabla_\theta u(\theta; Y)\}$
- variability matrix $J(\theta) = \mathrm{var}_\theta\{u(\theta; Y)\}$

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

# Inference

- Sample $y_1, \ldots, y_n$ independent from $f(y;\theta)$ or $f_i(y;\theta)$
- Composite log-likelihood $c\ell(\theta; y) = \sum_{i=1}^{n} c\ell(\theta; y_i)$
- maximum composite likelihood estimator
  $\hat{\theta}_{CL} = \arg\max c\ell(\theta; y)$ $u(\hat{\theta}_{CL}; y) = 0$
- Asymptotic consistency, normality

$$\sqrt{n}(\hat{\theta}_{CL} - \theta) \xrightarrow{\mathcal{L}} N_p\{0, G^{-1}(\theta)\}, \quad n \to \infty, \; m \text{ fixed}$$

- if $n$ fixed and $m \to \infty$, need assumptions on replication
- examples include time series and spatial data

decaying correlations

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

# ... inference

- $\theta = (\psi, \lambda), \quad \psi \in \mathbb{R}^q, \lambda \in \mathbb{R}^{p-q}$
- inference based on maximum likelihood estimator

$$\hat{\theta}_{CL} \dot{\sim} N_p\{\theta, G^{-1}(\hat{\theta}_{CL})\} \implies \hat{\psi}_{CL} \dot{\sim} N_q\{\psi, G^{\psi\psi}(\hat{\theta}_{CL})\}$$

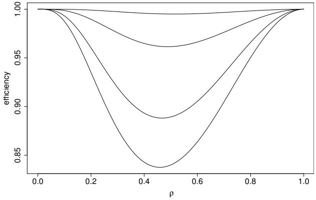
- CL log-likelihood ratio statistic
  $w_{CL}(\psi) = 2\{c\ell(\hat{\theta}_{CL}) - c\ell(\tilde{\theta}_{CL})\} \xrightarrow{\mathcal{L}} \sum_{j=1}^{q} \lambda_j Z_j^2$

- $\tilde{\theta}_{CL} = \{\psi, \hat{\lambda}_{CL}(\psi)\}$, constrained mle
- $Z_j \sim N(0, 1) \qquad \lambda_j$ eigenvalues of $(H^{\psi\psi})^{-1} G^{\psi\psi}$ Kent (1982)

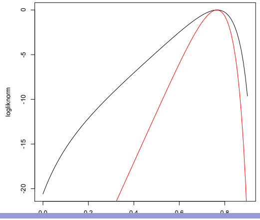
- $G(\theta) = H(\theta) J^{-1}(\theta) H(\theta)$

# ... inference

- $\hat{\theta}_{CL}$ not fully efficient unless $G(\theta) = H(\theta)J^{-1}(\theta)H(\theta) = i(\theta)$



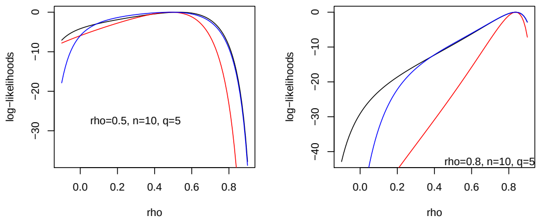

- $c\ell(\theta)$ is not a log-likelihood function



# ... inference

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

  Chandler & Bate (2007), Salvan et al. 11, 12 (wip)



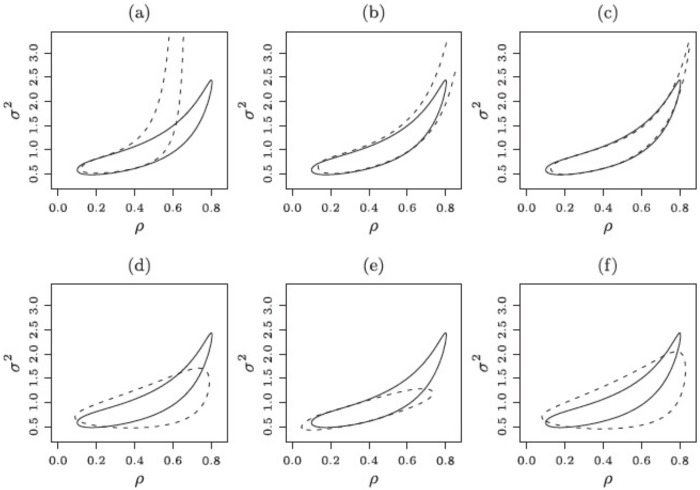
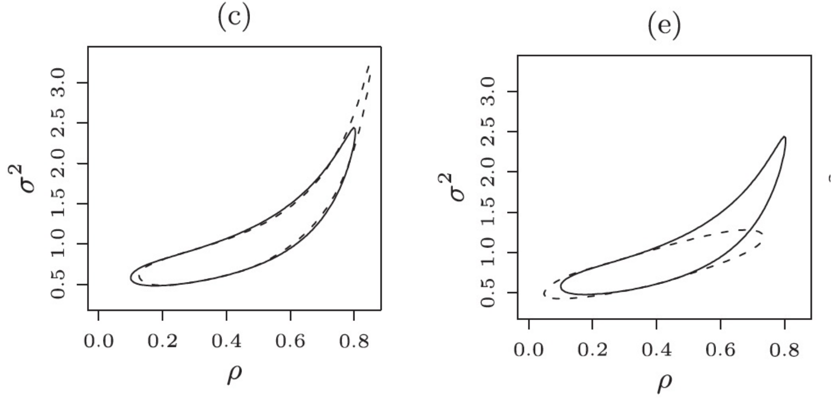



Figure 1. Equicorrelated multivariate normal model. Confidence regions with level 0.95 for $(\rho, \sigma^2)$, with $\mu$ known and equal to zero, for a simulated sample with $n = 5$, $q = 30$, and true parameter value $\mu = 0$, $\rho = 0.5$, and $\sigma^2 = 1$. In each plot, the solid line corresponds to $w(\theta)$, while the dashed line corresponds to: (a) $w^e(\theta)$; (b) $w^u(\theta)$; (c) $pw(\theta)_{INV}$; (d) $pw(\theta)_1$; (e) $pw(\theta)_{CB}$; (f) $pw(\theta)_2$.




$$\underline{y} \sim N(\mu\underline{1}, \sigma^2 R), \quad R_{rs} = \rho$$

# Model selection

- Akaike Information Criterion

$$AIC_{CL} = -2c\ell^n(\hat{\theta}_{CL}) + 2 \quad \mathrm{tr}\{J(\hat{\theta})H^{-1}(\hat{\theta})\}$$

Varin & Vidoni (2005)

- Bayesian Information Criterion

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

Gao & Song (2010)

- effective number of parameters

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

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

# Some surprises

- $\underline{Y} \sim N(\underline{\mu}, \Sigma) \quad \hat{\mu}_{CL} = \hat{\mu}, \hat{\Sigma}_{CL} = \hat{\Sigma}$
- $Y \sim N(\mu \underline{1}, \sigma^2 R), \quad R = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \ddots & \ddots & \vdots \\ \rho & \dots & \rho & 1 \end{pmatrix}$
  - $\hat{\theta}_{CL} = \hat{\theta}, \quad G(\theta) = i(\theta), \quad G(\theta) = H(\theta) J^{-1}(\theta) H(\theta)$
  - $H(\theta) \neq J(\theta) \qquad H(\theta) = \text{var(Score)}, \quad J = \text{E}(\nabla_\theta \text{Score})$
- $Y \sim (0, R)$: $\hat{\rho}_{CL} \neq \hat{\rho}$; a.var$(\hat{\rho}_{CL}) >$ a.var$(\hat{\rho})$
- efficiency improvement, nuisance parameter is unknown

  Mardia et al (2008); 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

# Applications: spatial and space-time data

- conditional approaches seem more natural
- condition on neighbours in space
- condition on 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, (2012); Genton et al., 12, Ribatet, 12

# Spatial extremes

- vector observations $(X_{1i}, \ldots, X_{di}), \; i = 1, \ldots, n$
- example, wind speed at each of $d$ 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

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

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

# ... 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 $V(\cdot)$
- 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

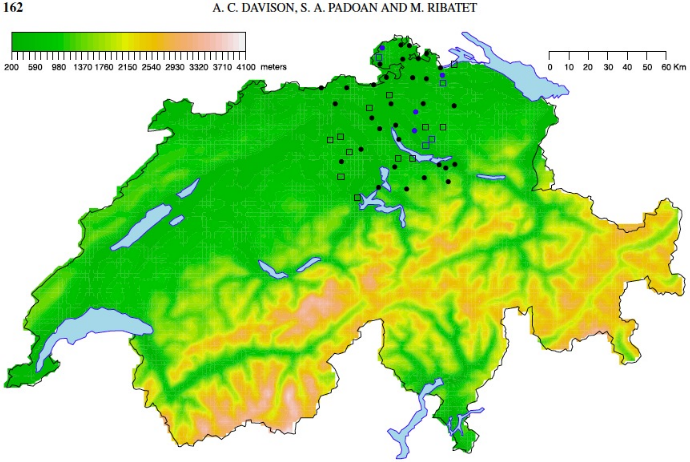



FIG. 1. *Map of Switzerland showing the stations of the* 51 *rainfall gauges used for the analysis, with an insert showing the altitude. The* 36 *stations marked by circles were used to fit the models, and those marked with squares were used to validate the models. Data for the pairs of stations with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

# ... Davison et al, 2012

MODELING OF SPATIAL EXTREMES 175

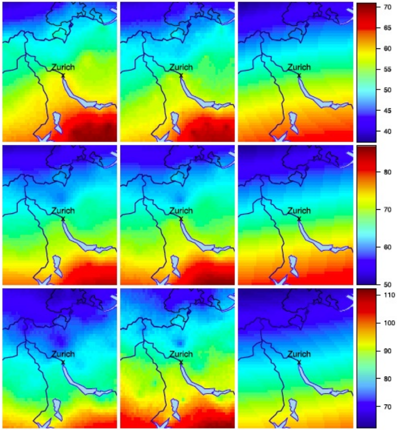
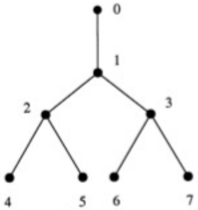



FIG. 3. *Maps of the (predictive) pointwise 25-year return level estimates for rainfall (mm) obtained from the latent variable and max-stable models. The top and bottom rows show the lower and upper bounds of the 95% pointwise credible/confidence intervals. The middle row shows the predictive pointwise posterior mean and pointwise estimates. The left column corresponds to the latent variable model assuming* Gamma(5, 3) *prior on* $\lambda$. *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.*

# Network tomography, Liang & Yu, 2003

- $X = (X_1, \ldots, X_{m'})$ network dynamics
  e.g. traffic flow counts; node delays
- $Y = (Y_1, \ldots, Y_m)$ measurement vector $m \ll m'$
- $Y = AX$, $\quad A$ known routing matrix, entries 1 $\quad m \times m'$
- components $X_j$ are independent, with density $f_j(\cdot; \theta_j)$



(a)

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_7 \end{pmatrix}$$

# ... 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 (2011)

- 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 (2011); 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 (IMS,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
- ...

# 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

# ... 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 Generalized Estimating Equations

# Aspects of robustness

- model robustness
  - univariate and bivariate margins only for example
  - means, variances, association parameters
  - similar in flavour to generalized estimating equations GEE: mean structure primary
- computational robustness
  - composite log-likelihood functions are smoother than log-likelihood functions
  - easier to maximize, easier to work with
  - especially in high dimension cases Liang and Yu (2003)
- robust to missing data mechanisms: Yi, Zeng and Cook (2010)
- access to multivariate distributions: e.g. mv extremes

  Davison et al. (2012)

# Robustness of consistency

working model $\{f(y;\theta);\theta \in \Theta\}$
true model $g(y)$

| Model | Full Likelihood | Composite Likelihood |
|---|---|---|
| Correctly specified | $f(y;\theta_0) = g(y)$ <br> $\hat{\theta}_{ML} \to \theta_0$ | $f_k(y;\theta_0) = g_k(y)$ for all $k$ <br> $\hat{\theta}_{CL} \to \theta_0$ |
| Misspecified | $f(y;\theta) \neq g(y)$, <br> $\hat{\theta}_{ML} \to \theta^*_{ML}$ | $f_k(y;\theta) \neq g_k(y)$ for some $k$ <br> $\hat{\theta}_{CL} \to \theta^*$ |

Top row – efficiency

Ximing Xu, U Toronto

# ... robustness of consistency

working model $\{f(y;\theta);\theta \in \Theta\}$
true model $g(y)$

| Model | Full Likelihood | Composite Likelihood |
|---|---|---|
| Correctly specified | $f(y;\theta_0) = g(y)$<br>$\hat{\theta}_{ML} \to \theta_0$ | $f_k(y;\theta_0) = g_k(y)$ for all $k$<br>$\hat{\theta}_{CL} \to \theta_0$ |
| Misspecified | $f(y;\theta) \neq g(y)$,<br>$\hat{\theta}_{ML} \to \theta^*_{ML}$ | $f_k(y;\theta) \neq g_k(y)$ for some $k$<br>$\hat{\theta}_{CL} \to \theta^*$ |

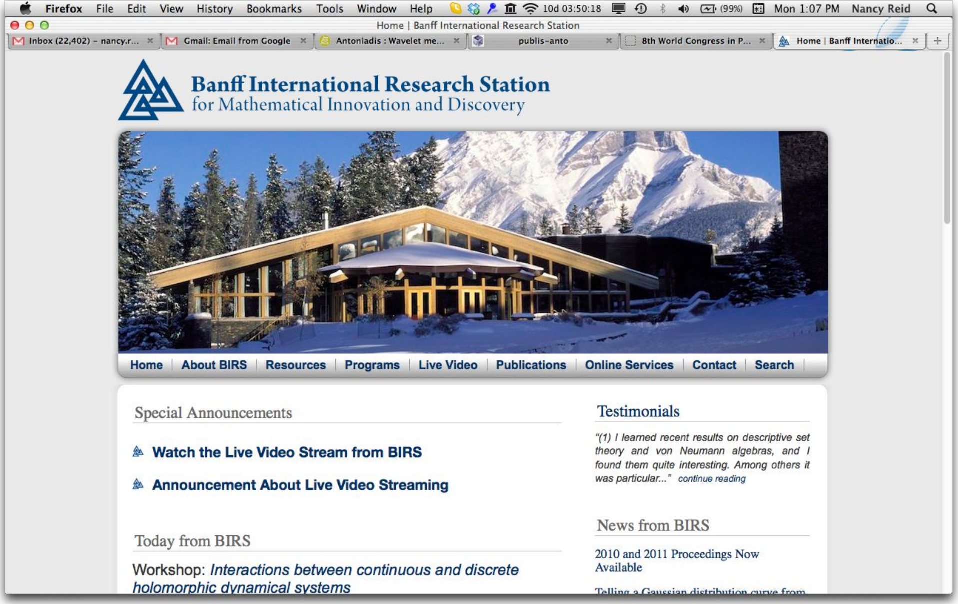
Ximing Xu, U Toronto

# ... robustness of consistency

- example (Andrei and Kendziorski, 2009): $Y_1 \sim N(\mu_1, \sigma_1^2),\ Y_2 \sim N(\mu_2, \sigma_2^2),\ \varepsilon \sim N(0, 1)$
- $Y_3 = Y_1 + Y_2 + bY_1 Y_2 + \epsilon$
- full likelihood for multivariate normal is a mis-specified model, $\hat{b} = 0$
- composite conditional likelihood based on normal distribution for $f(Y_3 \mid Y_2, Y_1) \longrightarrow$ consistent estimate of $b$
- example (Arnold and Xu): $f(Y) = \Phi_p(Y; \mu, \Sigma) + g(\mu, \Sigma)(\prod_{i=1}^{p} Y_i 1\{|Y_i| \leq t\})$
- sub-distributions of dimension $k < p$ are multivariate normal
- pairwise likelihood estimator nearly identical to mle assuming incorrect $N_p(\mu; \Sigma)$ model

# 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”
- ...


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