

# Composite likelihood methods

Nancy Reid

Texas A&M, May 13, 2008



## 60, 55, 30



[Emanuel Parzen, Bradley Efron, ... ... Bradley Efron, Emanuel Parzen] ...  
697 x 461 - 50k - jpg  
[www.stat.tamu.edu](http://www.stat.tamu.edu)



653 x 388 - 38k - jpg  
[www.stat.tamu.edu](http://www.stat.tamu.edu)



Ingram Olkin, Emanuel Parzen. ...  
626 x 475 - 46k - jpg  
[www.stat.tamu.edu](http://www.stat.tamu.edu)



Ingram Olkin, Emanuel Parzen. ...  
2505 x 1902 - 282k - jpg  
[www.stat.tamu.edu](http://www.stat.tamu.edu)



... and colleagues of Emanuel Parzen ...  
1494 x 967 - 110k - jpg  
[www.stat.tamu.edu](http://www.stat.tamu.edu)



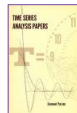
Emanuel Parzen  
177 x 266 - 36k - gif  
[stat.tamu.edu](http://stat.tamu.edu)



[Emanuel Parzen, Bradley Efron, ...  
348 x 237 - 18k - jpg  
[www.stat.tamu.edu](http://www.stat.tamu.edu)



Name: Emanuel Parzen  
918 x 1224 - 211k - jpg  
[www.stat.tamu.edu](http://www.stat.tamu.edu)



edited by Emanuel Parzen  
447 x 672 - 39k - jpg  
[www.stat.tamu.edu](http://www.stat.tamu.edu)



by Emanuel Parzen  
431 x 667 - 25k - jpg  
[www.stat.tamu.edu](http://www.stat.tamu.edu)



edited by Emanuel Parzen, ...  
475 x 719 - 69k - jpg  
[www.stat.tamu.edu](http://www.stat.tamu.edu)



The 2006 EMANUEL AND CAROL PARZEN ...  
773 x 528 - 252k - jpg  
[www.stat.tamu.edu](http://www.stat.tamu.edu)



The 1996 EMANUEL AND CAROL ... named after Emanuel Parzen) is  
a ...  
246 x 321 - 11k - jpg  
[www.stat.tamu.edu](http://www.stat.tamu.edu)



250 x 200 - 22k - png  
[en.wikipedia.org](http://en.wikipedia.org)



Author: Emanuel Parzen  
164 x 254 - 7k - jpg  
[images.bestwebbuys.com](http://images.bestwebbuys.com)

## 60, 55, 30

Your *continued donations* keep Wikipedia running!

[article](#) [discussion](#) [edit this page](#) [history](#)

[Log in / create account](#)

Early registration for Wikimania 2008 is now open. [Learn more about using Wikipedia for research](#)

## Kernel density estimation

From Wikipedia, the free encyclopedia

In statistics, **kernel density estimation** (or **Parzen window** method, named after Emanuel Parzen) is a way of estimating the probability density function of a random variable. As an illustration, given some data about a sample of a population, kernel density estimation makes it possible to *extrapolate* the data to the entire population.

A histogram can be thought of as a collection of point samples from a kernel density estimate for which the kernel is a uniform box the width of the histogram bin.

**Contents** [hide]

- 1 Definition
- 2 Intuition
- 3 Properties
- 4 Statistical implementation
- 5 See also
- 6 References
- 7 External links

Kernel density estimation of 100 normally distributed random numbers using different smoothing bandwidths.

[\[edit\]](#)

### Definition

[edit]



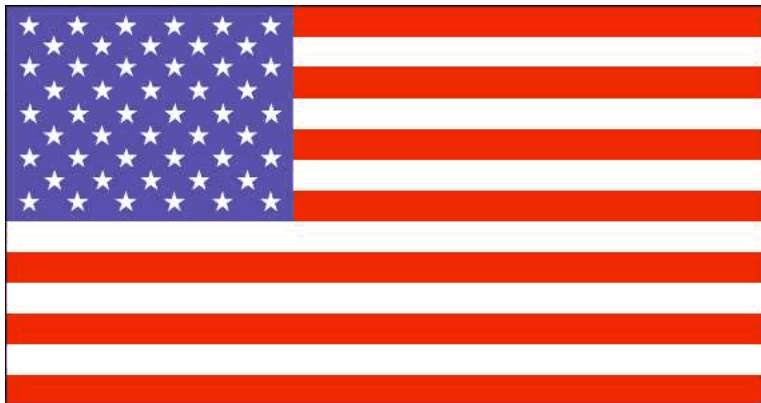
WIKIPEDIA  
The Free Encyclopedia

- navigation
- Main Page
  - Contents
  - Featured content
  - Current events
  - Random article

- interaction
- About Wikipedia
  - Community portal
  - Recent changes
  - Contact Wikipedia
  - Donate to Wikipedia
  - Help

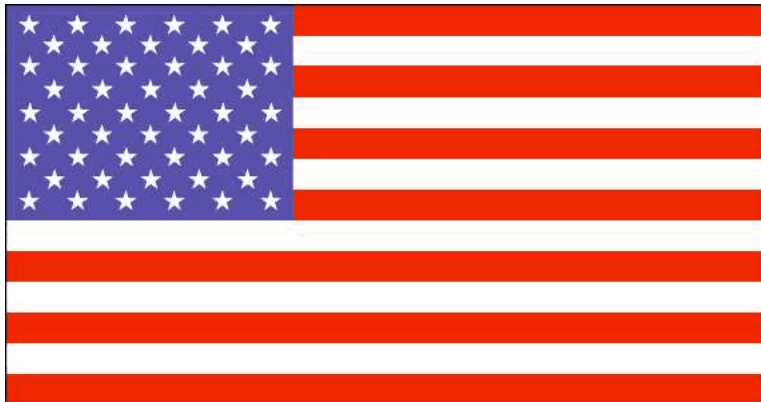
search

60, 55, 30



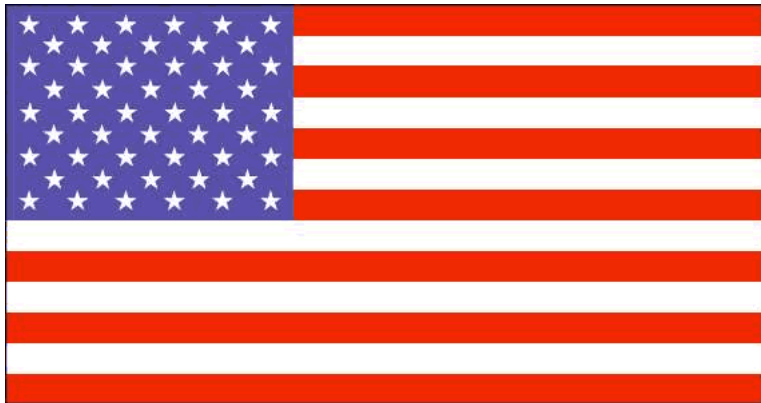
UAS:

# 60, 55, 30



UAS: Confidence, Philosophy, Teaching, History, Future, Careers, ...

# 60, 55, 30



UAS: Confidence, Philosophy, Teaching, History, Future, Careers, ...



## Composite likelihood

- ▶ **Model:**  $Y \sim f(y; \theta)$ ,  $Y \in \mathcal{Y} \subset \mathbb{R}^p$ ,  $\theta \in \mathbb{R}^d$
- ▶ **Likelihood function:**  $L(\theta; y) = f(y; \theta)$
- ▶ **Composite Marginal Likelihood:**  
 $CML(\theta; y) = \prod_{s \in \mathcal{S}} f_s(y_s; \theta)$ ,  $\mathcal{S}$  is a set of indices
- ▶ **Independence Likelihood:**  $\prod_{r=1}^p f_1(y_r; \theta)$
- ▶ **Pairwise Likelihood:**  $\prod_{r=1}^{p-1} \prod_{s=r+1}^p f_2(y_r, y_s; \theta)$
- ▶ **Composite Conditional Likelihood:** (Besag, 1974)  
 $CCL(\theta; y) = \prod_{s \in \mathcal{S}} f_{s|s^c}(y_s | y_{s^c})$ ,
- ▶ **Composite Likelihood:** (Lindsay, 1988)  $CL(\theta; y) = \prod_s L_s(\theta; y)$ ; each  $L_s(\theta; y)$  proportional to a conditional or marginal density

## Example: multi-level probit model

- ▶ binary observations with random effects:

$$y_{ir}, \quad r = 1, \dots, n_i; i = 1, \dots, N$$

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

- ▶ likelihood

$$L(\beta, \sigma_b) = \prod_{i=1}^N \log \int_{-\infty}^{\infty} \prod_{r=1}^{n_i} Pr(y_{ir} = 1 \mid b_i) \phi(b_i, \sigma_b^2) db_i$$

- ▶ 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  $\Phi_2(\cdot, \cdot; \rho_{irs})$
- ▶ using latent variable formulation:

$$z_{ir} = x'_{ir}\beta + b_i + \epsilon_{ir}, \quad \epsilon_{ir} \sim N(0, 1)$$

(Renard et al., 2004)

## ... multi-level probit (Renard et al. 2004)

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

## Derived quantities

- ▶ log composite likelihood:  $cl(\theta; y) = \log CL(\theta; y)$
- ▶ score function:  $U(\theta; y) = \nabla_{\theta} cl(\theta; y) = \sum_{s \in \mathcal{S}} w_s U_s(\theta; y)$
- ▶ maximum composite likelihood est.:  $\hat{\theta}_{CL} = \arg \sup cl(\theta; y)$
- ▶ variance:

$$J(\theta) = \text{var}_{\theta}\{U(\theta; Y)\}$$

$$H(\theta) = E_{\theta}\{-\nabla_{\theta} U(\theta; Y)\}$$

- ▶ Godambe information:

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

## Inference

▶ **Sample:**  $Y_1, \dots, Y_n$ , i.i.d.,  $CL(\theta; \underline{y}) = \prod_{i=1}^n CL(\theta; y_i)$

▶

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

▶  $w(\theta) = 2\{cl(\hat{\theta}_{CL}) - cl(\theta)\} \sim \sum_{a=1}^d \mu_a Z_a^2 \quad Z_a \sim N(0, 1)$

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

▶  $w(\psi) = 2\{cl(\hat{\theta}_{CL}) - cl(\tilde{\theta}_\psi)\} \sim \sum_{a=1}^{d_0} \mu_a Z_a^2$

▶ constrained estimator:  $\tilde{\theta}_\psi = \sup_{\theta=\theta(\psi)} cl(\theta; \underline{y})$

▶  $\mu_1, \dots, \mu_{d_0}$  eigenvalues of  $(H^{\psi\psi})^{-1}G^{\psi\psi}$

▶

Kent, 1982

## 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

$$cl(\rho; y_1, \dots, y_n) = -\frac{np(p-1)}{4} \log(1-\rho^2) - \frac{\rho-1+\rho}{2(1-\rho^2)} SS_w$$

$$- \frac{(\rho-1)(1-\rho)}{2(1-\rho^2)} \frac{SS_b}{\rho}$$

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

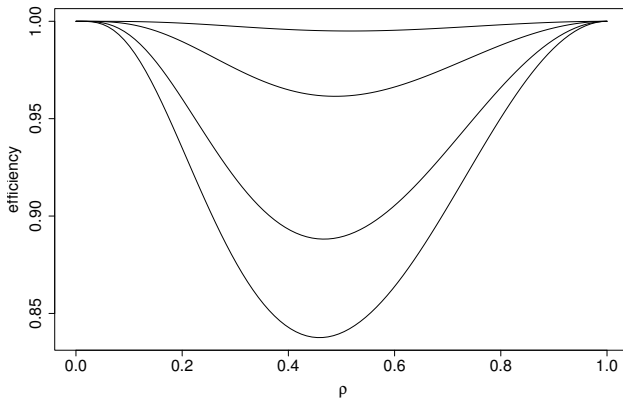
$$\ell(\rho; y_1, \dots, y_n) = -\frac{n(p-1)}{2} \log(1-\rho) - \frac{n}{2} \log\{1 + (p-1)\rho\}$$

$$- \frac{1}{2(1-\rho)} SS_w - \frac{1}{2\{1 + (p-1)\rho\}} \frac{SS_b}{\rho}$$

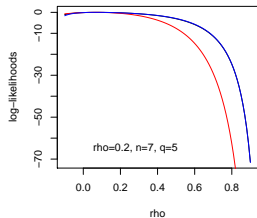
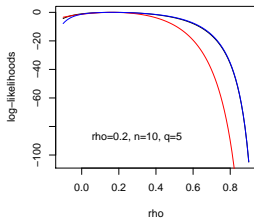
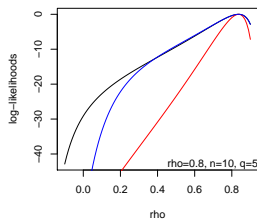
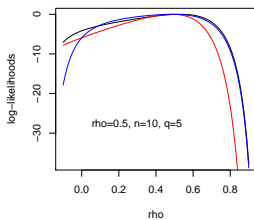
## ... symmetric normal

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

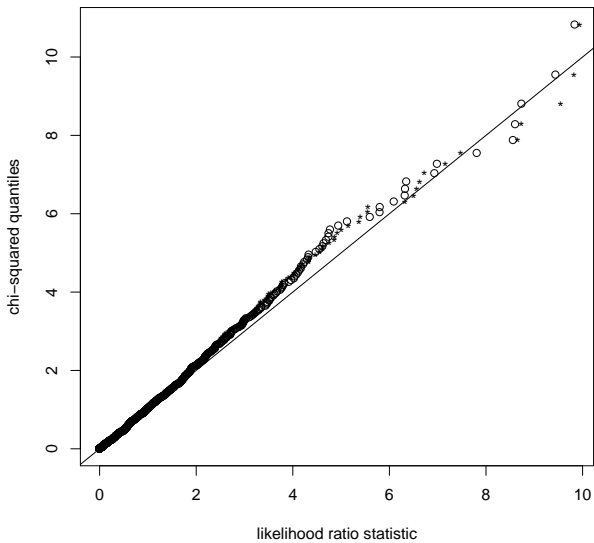
(Cox & Reid, 2004)



# Likelihood ratio test



$n=10, q=5, \rho=0.8$



\* – pairwise

## Motivation for composite likelihood

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

## Questions about modelling

- ▶ Does it matter if there is not a multivariate distribution compatible with, e.g., bivariate margins?
- ▶ Does theory of multivariate copulas help in understanding this?
- ▶ How do we ensure identifiability of parameters?
  - examples of trouble?
- ▶ Relationship to modelling via GEE?
- ▶ In what sense is it more robust?
- ▶ E.g. binary data using dichotomized MV Normal

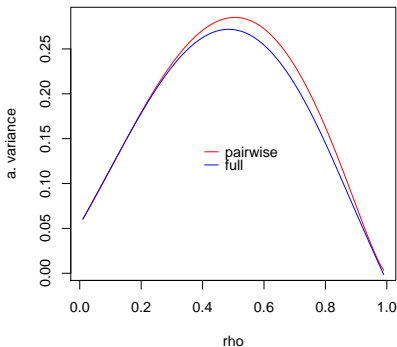
## Example: dichotomized MV Normal

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

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

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

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



$\rho$	0.02	0.05	0.12	0.20	0.40	0.50
ARE	0.998	0.995	0.992	0.968	0.953	0.968
$\rho$	0.60	0.70	0.80	0.90	0.95	0.98
ARE	0.953	0.903	0.900	0.874	0.869	0.850

## Questions about inference

- ▶ Efficiency of composite likelihood estimator:
  - ▶ choice of weights: Lindsay, 1988; Kuk and Nott, 2000;
  - ▶ assessment by simulation or direct comparison of a. var: Maydeu-Olivares and Joe, 2005
  - ▶ comparing two-stage to full pairwise estimation methods: Zhao and Joe, 2005; Kuk, 2007
  - ▶ ...
  
- ▶ Example: multivariate normal: Mardia, Hughes and Taylor (CCL), Jin (CML)
  - ▶  $Y \sim N(\underline{\mu}, \Sigma)$ : pairwise likelihood estimates  $\equiv$  mles
  - ▶  $Y \sim N(\underline{\mu}, \sigma^2 R)$ ,  $R_{ij} = \rho$ : pairwise likelihood est.  $\equiv$  mles
  - ▶  $Y \sim N(\underline{\mu}, R)$ : loss of efficiency (although small)
  
- ▶ ? Can we *explain* efficiency?
  - Hjort and Varin, 2005: Markov chains

## Questions about inference

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

$$p \rightarrow \infty$$

- ▶ single long time series
- ▶ spatial models ( $p$  indexes spatial sites)
- ▶ usually assume decaying correlations, so  $p$  can play the role of  $n$
- ▶ population genetics: estimation of the population recombination rate
- ▶ data is long sequence of alleles
- ▶ likelihood for each pair of segregating sites estimated by simulation
- ▶ pairwise likelihood formed by combining these
- ▶ Fearnhead & Donnelly, 2001; McVean et al., 2002; Fearnhead, 2003; Hudson, 2001

$$\dots p \rightarrow \infty$$

symmetric normal

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

$$\begin{array}{cc} O\left(\frac{1}{n}\right) & O(1) \\ n \rightarrow \infty & p \rightarrow \infty \end{array}$$

dichotomized mv normal:

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

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

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

## Continuous responses

- ▶ Multivariate Normal:

$$Y_i = (Y_{1i}, \dots, Y_{ki}) \sim N\{\beta_0 + \beta_1 x_i, \sigma^2 R_i(\alpha)\}$$

Zhao and Joe, 2005

- ▶ pairwise likelihood very efficient, but not  $\equiv$  max. lik. ARE
- ▶ multivariate longitudinal data; correlated series of observations with random effects
 

Fieuwis and Verbeke, 2006)
- ▶ correlation of full likelihood and pairwise likelihood estimates of parameters near 1, relative efficiency also near 1      simulations
- ▶ pairwise likelihood based on differences within clusters, and connections to within and between block analysis
 

Lele and Taper, 2002; Oakes and Ritz, 2000
- ▶ and several papers on survival data, often using copulas

## CL2

$\beta_0$	$\beta_1$	$\sigma^2$	$\rho$
0.998	0.997	1.000	0.913
0.996	0.995	1.000	0.889
0.995	0.996	0.999	0.876
1.000	0.999	1.000	0.884
0.960	0.968	0.987	0.967
0.974	0.970	0.993	0.964
0.978	0.969	0.992	0.928
0.986	0.977	0.993	0.903
0.942	0.958	0.961	0.957
0.944	0.949	0.961	0.952
0.949	0.945	0.966	0.922
0.964	0.939	0.966	0.898
0.924	0.966	0.934	0.943
0.926	0.947	0.937	0.940
0.943	0.932	0.949	0.925
0.982	0.913	0.976	0.919

## Binary data

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

## ... binary data

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

## Time series

- ▶ going back to proposal by Azzalini, 1983
- ▶  $n = 1$  of more interest, with long time series and possibly decaying correlations
- ▶ Markov chain models: Hjort and Varin, 2005
- ▶ Varin:

$$\prod_{t=m+1}^n \prod_{i=1}^m f_2(y_t, y_{t-i}; \theta)$$

- ▶ seems counterintuitive but seems to give good estimates
- ▶ state space models, population dynamics: Andrieu, 2008

## And more...

- ▶ spatial data: multivariate normal, generalized linear models, CML based on differences, CCL and modifications, network tomography, data on a lattice, point processes
- ▶ image analysis: Nott and Ryden, 1999
- ▶ Rasch model, Bradley-Terry model, ...
- ▶ space-time data
- ▶ block-based likelihoods for geostatistics: Caragea and Smith, 2007
- ▶ model selection using information criteria based on CL: Varin and Vidoni
- ▶ improvements of usual CL methods for specific models

## Questions

- ▶ is PL useful for modelling when no joint distribution is available (and may not exist?)
- ▶ e.g. extreme values, survival data (Parner, 2001)
- ▶ asymptotic theory in  $n, q$  together
- ▶ likelihood ratio type tests immediately available; one advantage over GEE
- ▶ can we really think beyond means and covariances in multivariate settings?
- ▶ should inference for mean parameters be separated from inference for covariances
- ▶ how to investigate robustness systematically
- ▶ estimation of Godambe information
- ▶ why does it work well? (when does it not work?)

# References

Varin, C. (2008) On composite marginal likelihoods. *Adv. Stat. Anal.* **95**, 1–28 [www.dst.unive.it/~sammy](http://www.dst.unive.it/~sammy)

Workshop on Composite Likelihood Methods, 15-17 April 2008



Workshop "survivors" on the last day (*bigger ; full-size*)

[www.utstat.utoronto.ca/reid/](http://www.utstat.utoronto.ca/reid/)

## ... References

- ▶ Lindsay, B. (1988) *Contemp. Math.* **80** 221–240
- ▶ Besag, J. (1974) *JRSS B* **34** 192–236
- ▶ Renard, D., Molenberghs, G. and Geys, H. (2004) *Comp. Stat. Data Anal.* **44** 629–667
- ▶ Kent, J. (1982) *Biometrika* **69** 19–27
- ▶ Cox, D.R. and Reid, N. (2004) *Biometrika* **91** 729–737
- ▶ ...

# Thank you!!

