The Pairwise Expectation Maximization Algorithm for Fitting Parameter-Driven Models

Xanthi Pedeli and Cristiano Varin
xanthi.pedeli@unive.it, cristiano.varin@unive.it

Ca' Foscari University of Venice

IWSM 2017
3-7 July 2017, Groningen

 This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 699980

Outline

Framework

Ordinary likelihood often too hard to evaluate or even specify for models with complex interdependencies

Usual problems

- large dense covariance matrices
- high-dimensional integrals
- normalizing constants
- nuisance components

For example, models with unobservables u

$$L(\theta; y) = \int f(y|u;\theta) f(u;\theta) du$$

Hard when the integral is high-dimensional like in spatio-temporal statistics

Composite likelihoods

Surrogates of intractable likelihoods in highly structured models

The inference function is a pseudolikelihood constructed from the combination of low-dimensional (marginal or conditional) component likelihoods

General setup

- Y m-dimensional vector random variable with probability density function $f(y;\theta)$, $\theta \in \Theta$
- $\{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$ set of marginal or conditional events
- $L_k(\theta; y) \propto f(y \in \mathcal{A}_k; \theta)$

A composite likelihood (Lindsay, 1988) is the weighted product

$$L_C(\theta; y) = \prod_{k=1}^{K} L_k(\theta; y)^{w_k}$$

with weights $w_k \geq 0$

Conditional and marginal components

Conditional

- Besag (1974, 1975) pseudolikelihood
  $L_C(\theta; y) = \prod_{i=1}^{m} f(y_i | \text{neighbours of } y_i; \theta)$
- full conditionals $L_C(\theta; y) = \prod_{i=1}^{m} f(y_i | y_{(-i)}; \theta)$
- pairwise conditionals $L_C(\theta; y) = \prod_{i=1}^{m} \prod_{j=1}^{m} f(y_i | y_j; \theta)$

Marginal

- independence likelihood $L_C(\theta; y) = \prod_{i=1}^{m} f(y_i; \theta)$
- pairwise likelihood $L_C(\theta; y) = \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} f(y_i, y_j; \theta)$
- tripletwise $L_C(\theta; y) = \prod_{i=1}^{m-2} \prod_{j=i+1}^{m-1} \prod_{k=j+1}^{m} f(y_i, y_j, y_k; \theta)$
- blockwise...

Key quantities

Composite log-likelihood $\ell_C(\theta; y) = \sum_{k=1}^{K} \ell_k(\theta; y) w_k$

Composite score $u(\theta; y) = \nabla_\theta \ell_C(\theta; y)$

(unbiased under standard regularity conditions on each likelihood component)

Maximum composite likelihood estimator $u(\hat{\theta}_C) = 0$

Sensitivity matrix $\mathsf{H}(\theta) = \mathsf{E}_\theta\{-\nabla_\theta u(\theta; Y)\}$

Variability matrix $\mathsf{J}(\theta) = \mathsf{Var}_\theta\{u(\theta; Y)\}$

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

Inference

Sample of i.i.d. observations $y_1, \ldots, y_n$ from $f(y; \theta)$ on $\mathbb{R}^m$

Asymptotic consistency and normality for $n \to \infty$ and m fixed

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

Sandwitch-type asymptotic variance

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

In the full likelihood case, $\mathsf{H} = \mathsf{J}$

Often difficult to estimate the variability matrix J (e.g. time series and spatial models)

Pairwise likelihood in linear time series models

Davis and Yau (2011)

Most of the dependence usually occurs in neighboring observations and decreases as the time lag between observations increases

Including all possible pairs of observations in the pairwise likelihood

$$L_P(\theta; y) = \prod_{i=j+1}^{n} \prod_{j=1}^{n-1} f(y_i, y_j; \theta)$$

can result in efficiency loss
Using pairs of observations up to a certain lag, say d,

$$L_P^{(d)}(\theta; y) = \prod_{i=d+1}^{n} \prod_{j=1}^{d} f(y_i, y_{i-j}; \theta)$$

performs much better

The corresponding maximum pairwise likelihood estimators (MPLE) of order d are consistent for both short- and long-memory processes

Pairwise likelihood in linear time series models

Davis and Yau (2011)

AR(1) model: MPLE of order 1 fully efficient since they coincide with the Yule-Walker estimators. Efficiency decreases with the inclusion of pairs at lag distance > 1

AR(p) model: MPLE of order p is the best choice (extension of the previous argument)

Pairwise likelihood in a parameter-driven time series model

GLM with latent autoregressive structure

$$\begin{aligned}
\text{latent process:} \quad & Z_t = x_t^T \boldsymbol{\beta} + U_t, \\
& U_t = \phi U_{t-1} + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} N(0, \sigma^2); \quad |\phi| < 1 \\
\text{observed process:} \quad & Y_t | Z_t \sim \text{GLM}(\mu_t, \gamma_t), \\
& \mu_t = g(Z_t), \ \text{var}(Y_t) = \gamma_t V(\mu_t)
\end{aligned}$$

Exact likelihood

$$L(\boldsymbol{\theta}; \boldsymbol{y}) \propto \int \ldots \int \prod_{t=1}^{n} f(y_t | z_t) f(z_1, \ldots, z_n; \boldsymbol{\theta}) dz_1 \ldots dz_n$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi, \sigma^2)$

Pairwise likelihood in a parameter-driven time series model

Direct computation of the exact likelihood is not feasible

Standard approach: simulation-based methods (e.g. MCMC, importance sampling, integrated nested Laplace approximation (INLA))

Alternative: pairwise likelihood estimation

$$L_p^{(d)}(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{t=d+1}^{n} \prod_{i=1}^{d} \int \int f(y_{t-i}|z_{t-i}) f(y_t|z_t) f(z_{t-i}, z_t; \boldsymbol{\theta}) dz_{t-i} dz_t$$

e.g. for $d = 1$, only consecutive pairs of observations are considered:

$$L_p^{(1)}(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{t=2}^{n} \int \int f(y_{t-1}|z_{t-1}) f(y_t|z_t) f(z_{t-1}, z_t; \boldsymbol{\theta}) dz_{t-1} dz_t$$

Pairwise likelihood in a parameter-driven time series model

For the computation of $L_p^{(1)}(\boldsymbol{\theta}|\boldsymbol{y})$, $n-1$ two-dimensional integrals need to be evaluated

Each of these integrals can be calculated by a *pairwise* EM algorithm with objective function

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)}) &= \sum_{t=2}^{n} \int\int \log f(y_{t-1}, y_t, z_{t-1}, z_t; \boldsymbol{\theta}) f(z_{t-1}, z_t | y_{t-1}, y_t; \boldsymbol{\theta}^{(i)}) dz_{t-1} dz_t \\
&= \sum_{t=2}^{n} \int\int \{\log f(y_{t-1}, y_t | z_{t-1}, z_t) + \log f(z_{t-1}, z_t; \boldsymbol{\theta})\} \\
&\qquad \times f(z_{t-1}, z_t | y_{t-1}, y_t; \boldsymbol{\theta}^{(i)}) dz_{t-1} dz_t
\end{aligned}
$$

Pairwise likelihood in a parameter-driven time series model

For the computation of $L_p^{(1)}(\boldsymbol{\theta}|\boldsymbol{y})$, $n-1$ two-dimensional integrals need to be evaluated

Each of these integrals can be calculated by a *pairwise* EM algorithm with objective function

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)}) &= \sum_{t=2}^{n} \int\int \log f(y_{t-1}, y_t, z_{t-1}, z_t; \boldsymbol{\theta}) f(z_{t-1}, z_t | y_{t-1}, y_t; \boldsymbol{\theta}^{(i)}) dz_{t-1} dz_t \\
&= \sum_{t=2}^{n} \int\int \{\log f(y_{t-1}, y_t | z_{t-1}, z_t) + \log f(z_{t-1}, z_t; \boldsymbol{\theta})\} \\
&\qquad \times f(z_{t-1}, z_t | y_{t-1}, y_t; \boldsymbol{\theta}^{(i)}) dz_{t-1} dz_t
\end{aligned}
$$

Pairwise likelihood in a parameter-driven time series model

For the computation of $L_p^{(1)}(\boldsymbol{\theta}|\boldsymbol{y})$, $n-1$ two-dimensional integrals need to be evaluated

Each of these integrals can be calculated by a *pairwise* EM algorithm with objective function

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)}) &= \sum_{t=2}^{n} \int\int \log f(y_{t-1}, y_t, z_{t-1}, z_t; \boldsymbol{\theta}) f(z_{t-1}, z_t | y_{t-1}, y_t; \boldsymbol{\theta}^{(i)}) dz_{t-1} dz_t \\
&= \sum_{t=2}^{n} \int\int \{\log f(y_{t-1}, y_t | z_{t-1}, z_t) + \log f(z_{t-1}, z_t; \boldsymbol{\theta})\} \\
&\qquad \times f(z_{t-1}, z_t | y_{t-1}, y_t; \boldsymbol{\theta}^{(i)}) dz_{t-1} dz_t \\
&\equiv \sum_{t=2}^{n} \int\int \log f(z_{t-1}, z_t; \boldsymbol{\theta}) f(z_{t-1}, z_t | y_{t-1}, y_t; \boldsymbol{\theta}^{(i)}) dz_{t-1} dz_t
\end{aligned}
$$

Pairwise likelihood in a parameter-driven time series model

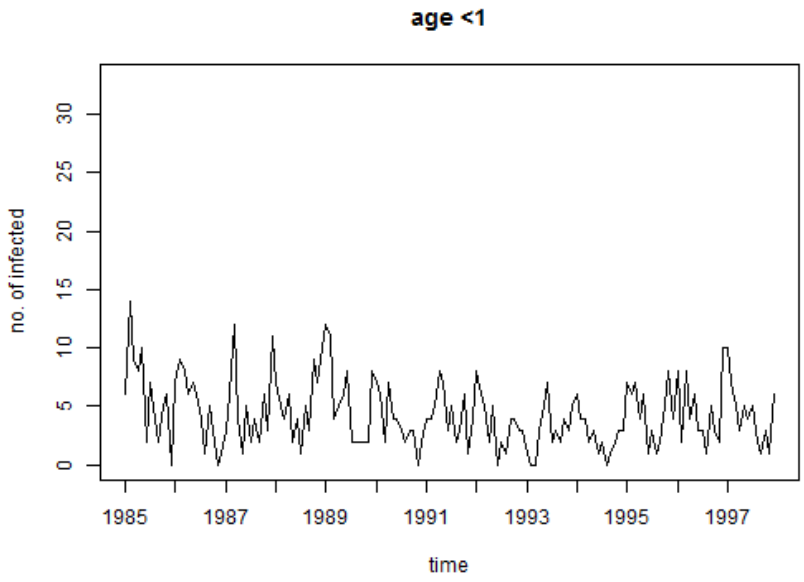
The unobserved process Z_t precludes evaluation of the double integrals involved in $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)})$ in closed form

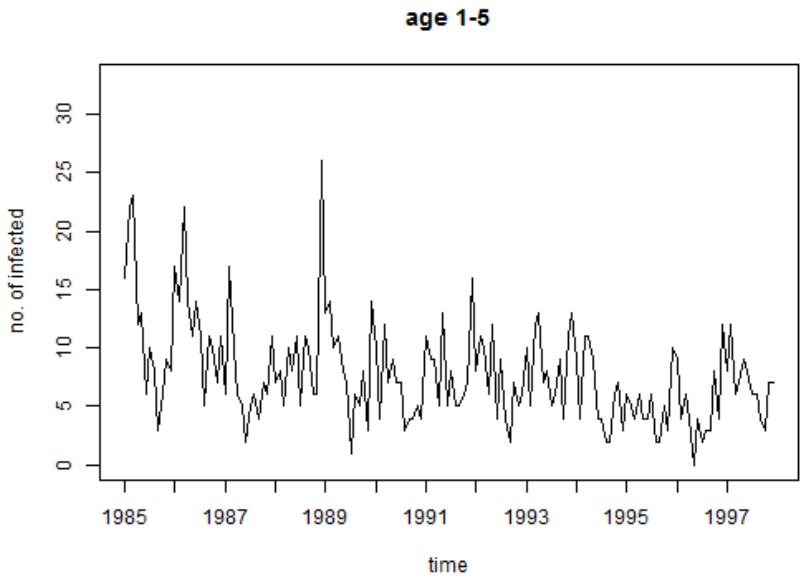
Gauss-Hermite quadrature: simple, efficient and deterministic approximation of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)})$ at the E-step of the pairwise EM

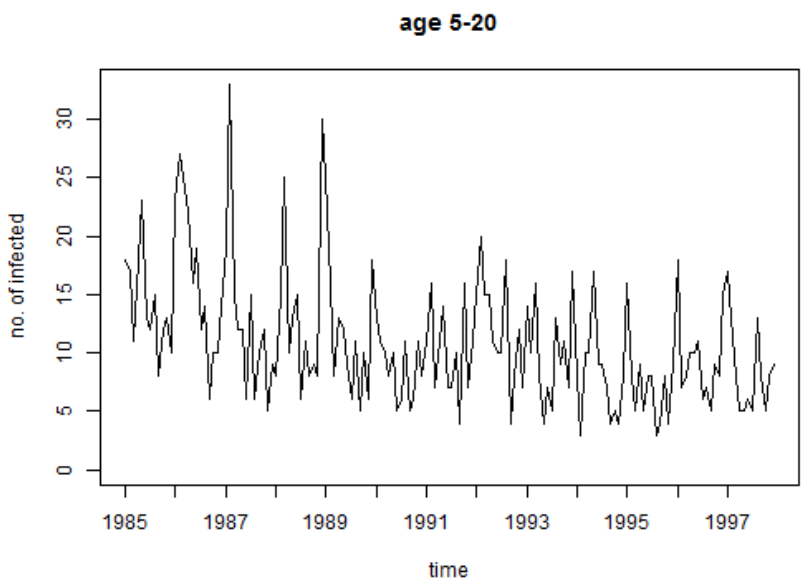
Further computational gain through a conditional maximization (CM) step: maximum pairwise likelihood estimators available in closed-form expressions

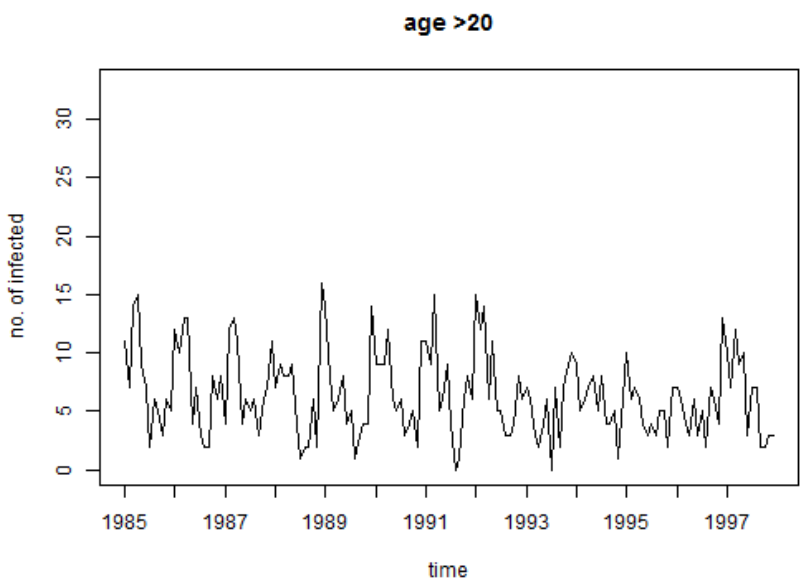
An application to health surveillance data

Monthly counts of meningococcal infections in France 1985-97 ($n = 156$)
(data available in the R package `surveillance`)









An application to health surveillance data

Analysis by age group accounting for
(i) trend and seasonality:

$$Y_t \sim \text{Pois}\left(\exp\left(\eta_t\right)\right),$$

where $\eta_t = \beta_0 + \beta_1 \cos\left(2\pi \frac{t}{12}\right) + \beta_2 \sin\left(2\pi \frac{t}{12}\right) + \beta_3 \frac{t}{156}$
(ii) trend, seasonality, autocorrelation:

$$Y_t | U_t \sim \text{Pois}\left(\exp\left(\eta_t + U_t\right)\right),$$

where $U_t = \phi U_{t-1} + \epsilon_t$, $\epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$

Starting values for the pairwise ECM algorithm:
- $\boldsymbol{\beta}^{(0)}$ obtained by model (i)
- $(\phi^{(0)}, \sigma^{(0)})$ taken by fitting an AR(1) model to the residuals of (i)

An application to health surveillance data

Poisson GLM

	age < 1	age $1-5$	age $5-20$	age > 20
cons.	1.639 (0.074)	2.371 (0.054)	2.706 (0.045)	1.954 (0.062)
cos	0.171 (0.055)	0.158 (0.041)	0.117 (0.035)	0.205 (0.046)
sin	0.365 (0.056)	0.310 (0.042)	0.256 (0.035)	0.427 (0.046)
trend	-0.428 (0.134)	-0.746 (0.101)	-0.708 (0.085)	-0.306 (0.110)

Poisson AR(1) model

	age < 1	age $1-5$	age $5-20$	age > 20
cons.	1.600 (0.132)	2.334 (0.067)	2.670 (0.068)	1.932 (0.069)
cos	0.164 (0.076)	0.154 (0.050)	0.111 (0.051)	0.207 (0.051)
sin	0.368 (0.068)	0.309 (0.055)	0.249 (0.053)	0.428 (0.052)
trend	-0.424 (0.245)	-0.722 (0.121)	-0.690 (0.108)	-0.281 (0.132)
AR(1)	0.725 (0.352)	0.320 (0.370)	0.211 (0.240)	0.480 (0.434)
stdev	0.188 (0.073)	0.211 (0.055)	0.228 (0.043)	0.133 (0.057)

Extension to the multidimensional space (work in progress)

The literature on multivariate time series models for count data is limited, partly due to the sharp increase in the complexity of maximum likelihood estimation

Ignoring cross correlation between series can result in misleading inference (e.g. correlation between age groups in the previous example)

Composite likelihood methods can facilitate estimation in the multidimensional space (Pedeli and Karlis, 2013) and provide a useful inferential tool in the definition of new multivariate time series models for counts

Extension to the multidimensional space (work in progress)

Example: GLM with a latent VAR(1) structure

Consider a bivariate latent process $\boldsymbol{Z}_t = \boldsymbol{X}_t^T \beta + \boldsymbol{U}_t$, where $\boldsymbol{X}_t = \begin{bmatrix} \boldsymbol{x}_{1t} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{x}_{2t} \end{bmatrix}$, $\beta = (\beta_1, \beta_2)^T$ and $\boldsymbol{U}_t$ is a bivariate VAR(1) process, i.e.

$$\boldsymbol{U}_t = \boldsymbol{\Phi} \boldsymbol{U}_{t-1} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{\Phi}$ (2×2) coefficient matrix and $\boldsymbol{\epsilon}_t$ (2×1) bivariate white noise vector with $E(\boldsymbol{\epsilon}_t) = 0$ and $E(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t^T) = \Sigma$

Given the latent process $\boldsymbol{Z}_t = (Z_{1t}, Z_{2t})^T$, each of the observed series $\boldsymbol{Y}_t = (Y_{1t}, Y_{2t})^T$ is

$$Y_{jt} | Z_{jt} \sim \text{GLM}(\mu_{jt}, \gamma_{jt}),$$

with $\mu_{jt} = g(Z_{jt})$, $\text{var}(Y_{jt}) = \gamma_{jt} V(\mu_{jt})$ for $j = 1, 2$

Extension to the multidimensional space (work in progress)

Exact likelihood includes a $(2n)$-dimensional integral:

$$L(\boldsymbol{\theta}; \boldsymbol{y}) \propto \int_{\mathbb{R}^n} \prod_{t=1}^{n} f(\boldsymbol{y}_t|\boldsymbol{z}_t) f(\boldsymbol{z}_t; \boldsymbol{\theta}) d\boldsymbol{z}_t$$

Pairwise likelihood of order d could considerably reduce the computational burden (only a certain number of 4-dimensional integrals need to be evaluated):

$$\begin{aligned} L_P^{(d)}(\boldsymbol{\theta}; \boldsymbol{y}) \quad = \quad & \prod_{t=d+1}^{n} \prod_{i=1}^{d} \int\int\int\int f(y_{1,t-i}|z_{1,t-i}) f(y_{2,t-i}|z_{2,t-i}) \\ & \times f(y_{1t}|z_{1t}) f(y_{2t}|z_{2t}) \\ & \times f(z_{1,t-i}, z_{2,t-i}, z_{1t}, z_{2t}; \boldsymbol{\theta}) dz_{1,t-i} dz_{2,t-i} dz_{1t} dz_{2t} \end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \text{vec}(\boldsymbol{\Phi}), \text{vec}(\boldsymbol{\Sigma}))$

Summary

- Likelihood-type inference in parameter-driven models for regression analysis of non-normal data in presence of serial correlation
- Exact maximum likelihood intractable but pairwise likelihood only requires to approximate a limited set of two-dimensional integrals
- Maximization of the pairwise likelihood through a pairwise version of the expectation maximization algorithm very convenient since estimators are available in closed-form expressions
- Promising for extensions to the multidimensional space

References


R.A. Davis and C.Y. Yau
Comments on pairwise likelihood in time series models
Statistica Sinica, 21: 255–277, 2011


B. Lindsay
Composite likelihood methods
Contemporary Mathematics, 80: 220–239, 1988


X. Pedeli and D. Karlis
On composite likelihood estimation of a multivariate INAR(1) model
J. Time Series Anal., 34: 206–220, 2013


C. Varin, N. Reid and D. Firth
An overview of composite likelihood methods
Statistica Sinica, 21: 5–42, 2011


X. Xu and N. Reid
On the robustness of maximum composite likelihood estimate
J. Stat. Plan. Inf., 141: 3047–3054, 2011