Composite likelihood methods for space (and space-time) covariance models

Carlo Gaetan

Dipartimento di Statistica
Università Ca' Foscari Venezia [1]

Rennes, 27-29 August 2008

[1] Joint work with M. Bevilacqua† ,J. Mateu‡, E. Porcu‡ († Università Ca' Foscari - Venezia, Italy, ‡ Universitat Jaume I, Castellón, Spain).

Outline of the talk

Geostatistical approach

Estimation methods

(Weighted) composite likelihood method

Model selection criterion

Conclusions

Geostatistical approach I

- $Z = \{Z(\mathbf{s}, t)\}$, spatio-temporal Random Fields (RFs), $\mathbf{s} \in \mathbb{R}^d$ is a spatial location, $t \in \mathbb{R}$ is a time point

$$\begin{aligned} Z(\mathbf{s}, t) &= \mu(\mathbf{s}, t) + \varepsilon(\mathbf{s}, t) \\ \text{data} &= \text{large scale} + \text{small scale} \end{aligned}$$

- Assumption: $\mu(\mathbf{s}, t)$ known ($\mu(\mathbf{s}, t) = 0$) and $\mathrm{E}[Z(\mathbf{s}, t)]^2 < \infty$.
- Space-time covariance function

$$\mathrm{cov}(Z(\mathbf{s}_1, t_1), Z(\mathbf{s}_2, t_2))$$

Geostatistical approach II

- Weakly stationarity

$$\mathrm{cov}(Z(\mathbf{s},t),Z(\mathbf{s}',t')) = C(\mathbf{s}-\mathbf{s}',t-t') = C(\mathbf{h},u)$$

($\mathbf{h}=\mathbf{s}-\mathbf{s}'$, spatial lag, $u=t-t'$ temporal lag).

- The (semi) variogram (under weak stationarity)

$$\frac{\mathrm{var}[Z(\mathbf{s},t)-Z(\mathbf{s}',t')]}{2} = \gamma(\mathbf{h},u) = C(\mathbf{0},0) - C(\mathbf{h},u)$$

$\mathbf{h}=\mathbf{s}-\mathbf{s}'$, $u=t-t'$

- Since a covariance function must be conditionally positive definite, practical estimation generally requires the selection of some parametric class of covariance and the corresponding estimation of these parameters.

$$\gamma(\mathbf{h},u;\theta) \Longleftrightarrow C(\mathbf{h},u;\theta)$$

WLS method (Cressie, 1985)

- Non parametric estimation of $\gamma(\mathbf{h}, u)$

$$\hat{\gamma}(\mathbf{h}, u) = \frac{1}{2|N(\mathbf{h}, u)|} \sum_{(\mathbf{s}_i, \mathbf{s}_j; t_i, t_j) \in N(\mathbf{h}, u)} (Z(\mathbf{s}_i, t_i) - Z(\mathbf{s}_j, t_j))^2$$

  where $N(\mathbf{h}, u)$ is some specified tolerance region around $\mathbf{h}$ and u (*bin*).

-

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \sum_{k=1}^{m} \frac{|N(\mathbf{h}_k, u_k)|}{\gamma^2(\mathbf{h}_k, u_k; \theta)} \left(\hat{\gamma}(\mathbf{h}_k, u_k) - \gamma(\mathbf{h}_k, u_k; \theta)\right)^2,$$

Maximum likelihood (ML) estimation

- Data: single realization $\mathbf{Z} = (Z(\mathbf{s}_1, t_1), \ldots, Z(\mathbf{s}_n, t_n))'$ from a space-time random field .
- $\{Z(\mathbf{s}, t)\}$ is zero mean Gaussian field. The log-likelihood

$$l(\theta) = -\frac{1}{2}\log \det\Sigma(\theta) - \frac{1}{2}\mathbf{Z}'\Sigma(\theta)^{-1}\mathbf{Z}$$

  where $\Sigma(\theta) = \mathrm{cov}(\mathbf{Z})$.
- Difficulties: for Gaussian random fields, the most critical part of the likelihood calculation is to evaluate the determinant and inverse of the covariance matrix. Each calculation of the likelihood requires $O(n^3)$ steps.

Composite likelihoods

General idea

1. Let $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ be a n-dimensional vector random variable with density $f(\mathbf{Z}; \theta)$ for some unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^d$.
2. Suppose that the joint distribution of Y is difficult to evaluate, but that it is possible to compute likelihoods for some subsets of the data.
3. It may be expedient to consider instead a pseudolikelihood compounding such likelihood objects.
4. This idea dates back to Besag (1974) and it has been termed composite likelihood after Lindsay (1988).

Composite likelihood: definition

Consider

1. a parametric model $\left\{f(\mathbf{Z};\theta), \mathbf{Z} \in \mathcal{Z} \subseteq \mathbb{R}^n, \theta \in \Theta \subseteq \mathbb{R}^p\right\}$;
2. a set of measurable events $\{\mathcal{A}_i; i = 1, \ldots, m\}$.

Then, a composite likelihood (CL) is the weighted product of the likelihoods corresponding to each single event,

$$\mathrm{CL}(\theta) = \mathrm{CL}(\theta;\mathbf{Z}) = \prod_{i=1}^{m} f(\mathbf{Z} \in \mathcal{A}_i;\theta)^{w_i},$$

where $\{w_i; i = 1, \ldots, m\}$ are positive weights.
Its maximum, if unique, is the **maximum composite likelihood estimator** (MCLE).

Vecchia (1988)'s approximation (spatial case)

- The exact joint density of $\mathbf{Z}$ may be written as

$$f(\mathbf{Z};\theta) = f(Z(s_1);\theta)\prod_{i=2}^{n} f(Z(s_i)|Z(s_{i-1}),\ldots,Z(s_1);\theta)$$

  where the ordering of observations is **arbitrary**.

- Replace

$$f(Z(s_i)|Z(s_{i-1}),\ldots,Z(s_1);\theta) \quad \text{by} \quad f(Z(s_i)|\mathbf{Z}(N_i);\theta),$$

  where $\mathbf{Z}(N_i)$ is some subset of $\{Z(s_{i-1}),\ldots,Z(s_1)\}$ and $|\mathbf{Z}(N_i)|$ is not too large.

$$CL(\theta) = \prod_{i=1}^{n} f(Z(s_i)|\mathbf{Z}(N_i);\theta)$$

- Each $\mathbf{Z}(N_i)$ consisted of a number of near neighbors of $Z(s_i)$, though the precise choice of $\mathbf{Z}(N_i)$ was **arbitrary**.

Stein et al. (2004)'s approximation

- It might be more efficient to do it in blocks, evaluating conditional densities of the form

$$f(Z(s_i), \dots, Z(s_{i+k}) | \mathbf{Z}(N_i); \theta)$$

- It is not necessarily best to choose $\mathbf{Z}(N_i)$ consisting only of near neighbours of the observation or observations whose conditional density is being evaluated.
- There is an extension to the space-time data for regular monitoring on time (Stein, 2005).

Caragea and Smith (2006)'s approximation

'Small blocks method':

- The observation locations are grouped into blocks N_i, $i = 1, \ldots, k$ of roughly the same size.
- For each block, compute the joint density of all observations in that block $f(\mathbf{Z}(N_i); \theta)$
- The small blocks likelihood is the product of joint densities for all the blocks, treating the blocks as if they were mutually independent.

$$CL(\theta) = \prod_{i=1}^{k} f(\mathbf{Z}(N_i); \theta)$$

- No extension to space-time data

Composite likelihood (Curriero and Lele, 1999) I

- We assume

$$U_{ij} = Z(\mathbf{s}_i, t_i) - Z(\mathbf{s}_j, t_j) \sim \mathcal{N}(0, 2\gamma_{ij}(\theta))$$

where $\gamma_{ij}(\theta) = \gamma(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j; \theta)$.

- First idea (marginal composition)

$$CL(\theta) = \prod_{j=1}^{n} \prod_{j>i}^{n} f(U_{ij}; \theta)$$

$$logCL(\theta) = \sum_{j=1}^{n} \sum_{j>i}^{n} \log f(U_{ij}; \theta) = \sum_{j=1}^{n} \sum_{j>i}^{n} l(U_{ij}; \theta)$$

where:

$$l(U_{ij}; \theta) = -\frac{1}{2} \log \gamma_{ij}(\theta) + \frac{U_{ij}^2}{2\gamma_{ij}(\theta)}.$$

Composite likelihood (Curriero and Lele, 1999) II

Features:

- Similar to WLS, but unlike WLS, it does not require any subjective choice of the lag bins.
- The number of operations requested is $O(n^2)$.
- To obtain estimates of θ we maximise the function $CL(\theta)$ or equivalently solve the estimating equation

$$v_{CL}(\theta) = \sum_{i=1}^{n} \sum_{j>i}^{n} \nabla l(U_{ij};\theta) = \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{\nabla \gamma_{ij}(\theta)}{\gamma_{ij}(\theta)} \left(1 - \frac{U_{ij}^2}{2\gamma_{ij}(\theta)}\right) = 0.$$

- Estimating unbiased equation, irrespectively of the distributional assumptions imposed on U_{ij}.

Optimal estimating equation

Second idea: optimal estimating equation

- If the fourth-order joint distributions of U_{ij} is known it would be possible to come up with an optimal way of combining the individual score $v_{CL}(\theta)$:

$$(\mathbb{E}\nabla v_{CL}(\theta))^{T}[\mathbb{C}\mathrm{ov}(v_{CL}(\theta))]^{-1}v_{CL}(\theta) = 0.$$

- The covariance matrix $\mathbb{C}\mathrm{ov}(v_{CL}(\theta))$ has dimension $n^2 \times n^2$, and its inversion is computationally prohibitive for large n.

Weighted composite likelihood

- Our idea: instead of searching optimal weights we consider

$$\mathrm{WCL}(\theta, \mathbf{d}) = \frac{1}{W_{n,\mathbf{d}}} \sum_{i}^{n} \sum_{j>i}^{n} l(U_{ij};\theta) w_{ij}(\mathbf{d}),$$

or

$$v(\theta, \mathbf{d}) = \frac{1}{W_{n,\mathbf{d}}} \sum_{i}^{n} \sum_{j>i}^{n} \nabla l(U_{ij};\theta) w_{ij}(\mathbf{d}) = 0,$$

where

$$w_{ij}(\mathbf{d}) = \begin{cases} 1 & \|\mathbf{s}_i - \mathbf{s}_j\| \le d_s, |t_i - t_j| \le d_t, \quad \mathbf{d} = (d_s, d_t)' \\ 0 & otherwise \end{cases}$$

and $W_{n,\mathbf{d}} = \sum_i^n \sum_{j>i}^n w_{ij}(\mathbf{d})$.

- We look for an “optimal lag” $\mathbf{d}^*$.

A measure of efficiency

How to choose **d** ? We look at the Godambe information matrix

$$G(\theta, \mathbf{d}) = H(\theta, \mathbf{d}) J(\theta, \mathbf{d})^{-1} H(\theta, \mathbf{d})',$$

where

$$H(\theta, \mathbf{d}) = \mathbb{E}[\nabla v(\theta, \mathbf{d})]$$

and

$$J(\theta, \mathbf{d}) = \mathbb{E}[v(\theta, \mathbf{d}) v(\theta, \mathbf{d})']$$

In our case

$$H(\theta, \mathbf{d}) = \mathbb{E}[\nabla e_{\mathrm{WCL}}(\theta, \mathbf{d})] = \frac{1}{W_{n,\mathbf{d}}} \sum_i \sum_{j>i} \left(\frac{\nabla \gamma_{ij}}{\gamma_{ij}} \frac{\nabla \gamma_{ij}'}{\gamma_{ij}} \right) w_{ij}(\mathbf{d})$$

$$J(\theta, \mathbf{d}) = \mathbb{E}[e_{\mathrm{WCL}}(\theta, \mathbf{d}) e_{\mathrm{WCL}}(\theta, \mathbf{d})'] = \frac{2}{W_{n,\mathbf{d}}^2} \sum_i \sum_{j>i} \sum_l \sum_{k>l} \frac{\nabla \gamma_{ij}'}{\gamma_{ij}} \frac{\nabla \gamma_{lk}'}{\gamma_{lk}} \rho_{ijlk} w_{ij}(\mathbf{d}) w_{lk}(\mathbf{d})$$

where $\rho_{ijlk} = Corr(U_{ij}^2, U_{lk}^2)$
Under Gaussianity:

$$\rho_{ijlk} = Corr(U_{ij}^2, U_{lk}^2) = \frac{(\gamma_{il} - \gamma_{jl} - \gamma_{jk} + \gamma_{ik})^2}{4\gamma_{ij}\gamma_{lk}} \tag{1}$$

Asymptotics I

We suppose

- $\theta \in \Theta \subset \mathbb{R}^p$, Θ compact set;
- increasing domain asymptotics $R_0 = (-\frac{1}{2}, \frac{1}{2}]^{d+1}$, $R_n = \{(\mathbf{s}_1, t_1), \dots, (\mathbf{s}_n, t_n)\} = (nR_0) \cap \mathbb{Z}^{d+1}$
- $\mathcal{M} = \{(\mathbf{h}_1, u_1), \dots, (\mathbf{h}_K, u_K)\}$, $K \geq p$ is a finite set not containing the origin and which determines which pairs of observations contribute to the sum;
- $\gamma(h; \theta)$ is twice continuously differentiable for $\theta \in V$, V is a neighbourhood of the true value θ_0;
- $\Gamma(\theta) = [\nabla\gamma(\mathbf{h}_1, u_1; \theta), \dots, \nabla\gamma(\mathbf{h}_K, u_K; \theta)]$ has full rank
- $\sum_{i=1}^K (2\gamma(\mathbf{h}_i, u_i; \theta_1) - 2\gamma(\mathbf{h}_i, u_i; \theta_2)) > 0$ for all $\theta_1 \neq \theta_2$, (identifiability condition);
- a mixing conditions on $\{Z(s, t)\}$

Asymptotics II

then (Guyon, 1995)

- $-WCL(\theta)$ is an additive contrast function;
- $\hat{\theta}_{WCL}$ is consistent and asymptotically Gaussian

$$G(\theta, \mathbf{d})^{1/2}(\hat{\theta}_{WCL} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_p)$$

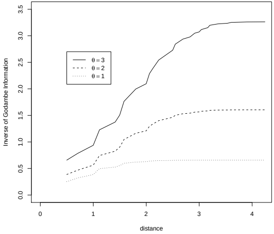
i.e.

$$(\hat{\theta}_{WCL} - \theta_0) \approx \mathcal{N}(0, G(\theta, \mathbf{d})^{-1})$$

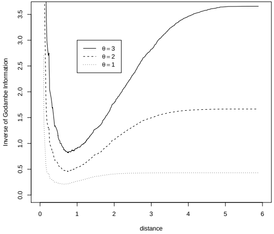
A simple spatial example

1. Exponential model

$$C(\mathbf{h};\theta) = \exp\left(-3\frac{\|\mathbf{h}\|}{\theta}\right), \quad \theta > 0. \qquad (2)$$



(a)



(b)

(a) 49 points located on a 7×7 regular grid $[0, 0.5, \ldots, 3]^2$;
(b) 49 points uniformly distributed on $[0, 3]^2$.

Weighted composite likelihood: practical implementation

- First step:
  We choose the 'lag' $\mathbf{d}$ minimising the $G^{-1}(\theta, \mathbf{d})$ in the partial order of nonnegative definite matrices or equivalently

$$\mathbf{d}^* = \operatorname*{argmin}_{\mathbf{d} \in \mathcal{D}} \operatorname{tr}(G^{-1}(\theta, \mathbf{d})), \qquad (3)$$

  where $\mathcal{D}$ is a set of lags.
    - Get a consistent estimate for θ (for instance $\hat{\theta}_{WLS}$)
    - Computation of $J(\hat{\theta}_{WLS}, \mathbf{d})$ becomes quickly infeasible ($O(n^4)$). Estimation through sub-sampling technique.
- Second step:

$$\hat{\theta}_{WCL} = \operatorname*{argmin}_{\theta \in \Theta} WCL(\theta, \mathbf{d}^*) \qquad (4)$$

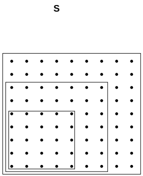
Computational burden

Method	Complexity	Drawbacks
Likelihood	$O(n^3)$	unfeasible for large data-set
Vecchia & Stein	$O(n)$	subjective conditional sets choice
Caragea & Smith	$O(n^2)$	subjective size of the block
WCLIC	$O(W^2_{n,\mathbf{d}^*})$	a preliminary estimation

A space-time example I

300 independent simulations from a zero mean Gaussian process on

- a space-time lattice $\mathcal{S} \times \mathcal{T}$, with
  - $\mathcal{S} = \{1, 1.5, 2, \ldots, N\}^2$ and $N = 3, 4, 5$



  - $\mathcal{T} = \{1, \ldots, T\}$ and $T = 15, 30, 45$

A space-time example II

- a non separable covariance model:

$$\mathrm{C}(\mathbf{h}, u) = \frac{1}{(a|u|+1)} \exp\left(-\frac{c\|\mathbf{h}\|}{(a|u|+1)^{0.25}} \right), \qquad a = c = 2$$

MSE Relative efficiency for WLS, CL and WCL estimation methods with respect to ML.

		$n = 25$			$n = 49$			$n = 81$		
		WLS	CL	WCL	WLS	CL	WCL	WLS	CL	WCL
	c	12.96	16.33	4.17	21.83	28.69	4.39	28.51	37.61	7.10
$T = 15$	a	12.85	21.59	4.38	20.35	30.53	4.92	25.19	33.71	6.95
	c	19.23	23.95	4.32	26.01	32.59	6.79	33.53	41.81	7.30
$T = 30$	a	25.34	40.91	5.23	33.39	46.59	5.79	39.91	49.95	7.23
	c	20.25	25.05	4.85	33.04	41.54	6.56	39.10	47.97	8.01
$T = 45$	a	27.34	41.83	4.35	40.96	51.12	5.18	46.86	58.53	6.42

Model selection criterion

- Model selection criteria as AIC and BIC depend on the computation of the likelihood function.
- We follow (Varin and Vidoni, 2005) and we select the model maximizing

$$WCLIC(\hat{\theta}_{\mathrm{WCL}}) = WCL(\hat{\theta}_{\mathrm{WCL}}) + \mathrm{tr}(\widehat{J}\widehat{H}^{-1}), \tag{5}$$

where $\widehat{J}$ and $\hat{H}$ are consistent estimates of J and H.

- If WCL $=$ L the the selection statistic reduces to the Akaike criterion

$$l(\hat{\theta}_{\mathrm{ML}}) - dim(\theta)$$

WCLIC: a simulation study

- 100 independent simulations from a zero mean space-time gaussian process with covariance models:

$$\mathrm{C}(\mathbf{h}, u) = \frac{\sigma^2}{(a|u|^{2\alpha}+1)} \exp\left(-\frac{c\|\mathbf{h}-\varepsilon u \mathbf{v}\|^{2\gamma})}{(a|u|^{2\alpha}+1)^{\beta\gamma}}\right). \qquad (6)$$

  1. A –Separable model ($\beta = 0, \varepsilon = 0$)
  2. B –Non separable model ($\varepsilon = 0$)
  3. C –Asymmetric in time non separable model

- $\mathcal{S}$ regular spaced grid on a square $[1,4]^2$ equally spaced by 1 (i.e. 16 locations) and $\mathcal{T} = \{1, \ldots, 150\}$

		Identified		
		A	B	C
	A	81	14	5
True	B	6	80	14
	C	3	11	86

Conclusions

- WCL seems to be a valid compromise between the computational burdens of ML and the loss of efficiency of WLS.
- Godambe information as natural criteria for the optimal distance for the WCL.
- Model selection is feasible for WCL.

Merci !

References I

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society B*, **36**, 192–236.

Caragea, P. and Smith, R. (2006) Approximate likelihoods for spatial processes. *Tech. rep.*, Department of Statistics, Iowa State University.

Cressie, N. (1985) Fitting variogram models by weighted least squares. *Mathematical Geology*, **17**, 239–252.

Curriero, F. and Lele, S. (1999) A composite likelihood approach to semivariogram estimation. *Journal of Agricultural, Biological and Environmental Statistics*, **4**, 9–28.

Guyon, X. (1995) *Random Fields on a Network: Modeling, Statistics and Applications*. New York: Springer.

Lindsay, B. (1988) Composite likelihood methods. *Contemporary Mathematics*, **80**, 221–239.

References II

Stein, M. (2005) Statistical methods for regular monitoring data. *Journal of the Royal Statistical Society B*, **67**, 667–687.

Stein, M., Chi, Z. and Welty, L. (2004) Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society B*, **66**, 275–296.

Varin, C. and Vidoni, P. (2005) A note on composite likelihood inference and model selection. *Biometrika*, **52**, 519–528.

Vecchia, A. (1988) Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society B*, **50**, 297–312.