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

Carlo Gaetan

Dipartimento di Statistica
Università Ca’ Foscari Venezia

Rennes, 27-29 August 2008

1Joint 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

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

$$Z(s, t) = \mu(s, t) + \varepsilon(s, t)$$

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

$$\text{cov}(Z(s_1, t_1), Z(s_2, t_2))$$

data = large scale + small scale
Geostatistical approach II

• Weakly stationarity

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

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

• The (semi) variogram (under weak stationarity)

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

$$h = s - s', u = t - t'$$

• 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.

$$\gamma(h, u; \theta) \iff C(h, u; \theta)$$
WLS method (Cressie, 1985)

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

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

where $N(h, u)$ is some specified tolerance region around $h$ and $u$ (bin).

- $$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{k=1}^{m} \frac{|N(h_k, u_k)|}{\gamma^2(h_k, u_k; \theta)} (\hat{\gamma}(h_k, u_k) - \gamma(h_k, u_k; \theta))^2,$$
Maximum likelihood (ML) estimation

- Data: single realization $Z = (Z(s_1, t_1), \ldots, Z(s_n, t_n))'$ from a space-time random field.
- $\{Z(s, t)\}$ is zero mean Gaussian field. The log-likelihood
  \[
  l(\theta) = \frac{-1}{2} \log \det \Sigma(\theta) - \frac{1}{2} Z' \Sigma(\theta)^{-1} Z
  \]
  where $\Sigma(\theta) = \text{cov}(Z)$.
- 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 \( \{ f(Z; \theta), Z \in \mathcal{Z} \subseteq \mathbb{R}^n, \theta \in \Theta \subseteq \mathbb{R}^p \} \);
2. a set of measurable events \( \{ A_i; i = 1, \ldots, m \} \).

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

\[
CL(\theta) = CL(\theta; Z) = \prod_{i=1}^{m} f(Z \in A_i; \theta)^{w_i},
\]

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
  \[
  f(Z(s_i)|Z(s_{i-1}), \ldots, Z(s_1); \theta)
  \]
  by
  \[
  f(Z(s_i)|\mathbf{Z}(N_i); \theta),
  \]
  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.

- 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

\[ f(Z(s_i), \ldots, Z(s_{i+k})|Z(N_i); \theta) \]

• It is not necessarily best to choose \( Z(N_i) \) 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 $f(Z(N_i); \theta)$
- The small blocks likelihood is the product of joint densities for all the blocks, treating the blocks as if they were mutually independent.

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

- No extension to space-time data
Composite likelihood (Curriero and Lele, 1999)

- We assume

\[ U_{ij} = Z(s_i, t_i) - Z(s_j, t_j) \sim \mathcal{N}(0, 2\gamma_{ij}(\theta)) \]

where \( \gamma_{ij}(\theta) = \gamma(s_i - s_j, t_i - t_j; \theta) \).

- First idea (marginal composition)

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

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

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 $\theta$ we maximise the function $CL(\theta)$ or equivalently solve the estimating equation

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

• 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 $\nu_{CL}(\theta)$:

$$
(\mathbb{E} \nabla \nu_{CL}(\theta))^T [\text{Cov}(\nu_{CL}(\theta))]^{-1} \nu_{CL}(\theta) = 0.
$$

- The covariance matrix $\text{Cov}(\nu_{CL}(\theta))$ 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

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

or

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

where

\[ w_{ij}(d) = \begin{cases} 
1 & \|s_i - s_j\| \leq d_s, |t_i - t_j| \leq d_t, \quad d = (d_s, d_t)' \\
0 & \text{otherwise} \end{cases} \]

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

- We look for an “optimal lag” \( d^*. \)
A measure of efficiency

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

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

where

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

and

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

In our case

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

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

where $\rho_{ijkl} = \text{Corr}(U_{ij}^2, U_{lk}^2)$

Under Gaussianity:

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

(1)
We suppose

- $\theta \in \Theta \subset \mathbb{R}^p$, $\Theta$ compact set;
- increasing domain asymptotics $R_0 = (-\frac{1}{2}, \frac{1}{2}]^{d+1}$, $R_n = \{(s_1, t_1), \ldots, (s_n, t_n)\} = (nR_0) \cap \mathbb{Z}^{d+1}$
- $\mathcal{M} = \{(h_1, u_1), \ldots, (h_K, u_K)\}$, $K \geq p$ 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 $\theta_0$;
- $\Gamma(\theta) = [\nabla \gamma(h_1, u_1; \theta), \ldots, \nabla \gamma(h_K, u_K; \theta)]$ has full rank
- $\sum_{i=1}^{K} (2\gamma(h_i, u_i; \theta_1) - 2\gamma(h_i, u_i; \theta_2)) > 0$ for all $\theta_1 \neq \theta_2$, (identifiability condition);
- a mixing conditions on $\{Z(s, t)\}$
then (Guyon, 1995)

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

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

i.e.

\[
(\hat{\theta}_{WCL} - \theta_0) \approx \mathcal{N}(0, G(\theta, d)^{-1})
\]
A simple spatial example

1. Exponential model

\[ C(h; \theta) = \exp\left( -3 \frac{\|h\|}{\theta} \right), \quad \theta > 0. \tag{2} \]

(a) 49 points located on a $7 \times 7$ 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’ $d$ minimising the $G^{-1}(\theta, d)$ in the partial order of nonnegative definite matrices or equivalently

$$d^* = \arg\min_{d \in D} \text{tr}(G^{-1}(\theta, d)),$$

where $D$ is a set of lags.

  ▶ Get a consistent estimate for $\theta$ (for instance $\hat{\theta}_{WLS}$)
  ▶ Computation of $J(\hat{\theta}_{WLS}, d)$ becomes quickly infeasible ($O(n^4)$). Estimation through sub-sampling technique.

• Second step:

$$\hat{\theta}_{WCL} = \arg\min_{\theta \in \Theta} WCL(\theta, d^*)$$
## Computational burden

<table>
<thead>
<tr>
<th>Method</th>
<th>Complexity</th>
<th>Drawbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood</td>
<td>$O(n^3)$</td>
<td>unfeasible for large data-set</td>
</tr>
<tr>
<td>Vecchia &amp; Stein</td>
<td>$O(n)$</td>
<td>subjective conditional sets choice</td>
</tr>
<tr>
<td>Caragea &amp; Smith</td>
<td>$O(n^2)$</td>
<td>subjective size of the block</td>
</tr>
<tr>
<td>WCLIC</td>
<td>$O(W_{n,d}^2)$</td>
<td>a preliminary estimation</td>
</tr>
</tbody>
</table>
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:

\[ C(h, u) = \frac{1}{(a|u| + 1)} \exp \left( -\frac{c\|h\|}{(a|u| + 1)^{0.25}} \right), \quad a = c = 2 \]

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

<table>
<thead>
<tr>
<th></th>
<th>( n = 25 )</th>
<th>( n = 49 )</th>
<th>( n = 81 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WLS</td>
<td>CL</td>
<td>WCL</td>
</tr>
<tr>
<td>( T = 15 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( c )</td>
<td>12.96</td>
<td>16.33</td>
<td>4.17</td>
</tr>
<tr>
<td>( a )</td>
<td>12.85</td>
<td>21.59</td>
<td>4.38</td>
</tr>
<tr>
<td>( T = 30 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( c )</td>
<td>19.23</td>
<td>23.95</td>
<td>4.32</td>
</tr>
<tr>
<td>( a )</td>
<td>25.34</td>
<td>40.91</td>
<td>5.23</td>
</tr>
<tr>
<td>( T = 45 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( c )</td>
<td>20.25</td>
<td>25.05</td>
<td>4.85</td>
</tr>
<tr>
<td>( a )</td>
<td>27.34</td>
<td>41.83</td>
<td>4.35</td>
</tr>
</tbody>
</table>
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

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

where $\hat{J}$ and $\hat{H}$ are consistent estimates of $J$ and $H$.
- If WCL = L the selection statistic reduces to the Akaike criterion

$$l(\hat{\theta}_{ML}) - \text{dim}(\theta)$$
WCLIC: a simulation study

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

\[
C(h, u) = \frac{\sigma^2}{(a|u|^{2\alpha} + 1)} \exp \left( -\frac{c \|h - \varepsilon uv\|^{2\gamma}}{(a|u|^{2\alpha} + 1)^{\beta \gamma}} \right).
\]  

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

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

<table>
<thead>
<tr>
<th>Identified</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>81</td>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>True</td>
<td>B</td>
<td>6</td>
<td>80</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>11</td>
<td>86</td>
</tr>
</tbody>
</table>
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


References II


