A restricted composite likelihood approach to modelling Gaussian geostatistical data

Kathy Mutambanengwe\textsuperscript{1,2}; Christel Faes\textsuperscript{2}; Marc Aerts\textsuperscript{2}

\textsuperscript{1}Open Analytics NV, Belgium; \textsuperscript{2}Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Universiteit Hasselt, Belgium

useR! 2017, Brussels, Belgium
July 5, 2017
Background

Methodology

Simulation study

spatialRECL package

Application to data
Background
Background

- Likelihood methods instinctive choice in modelling
- Not so straightforward with complex dependencies
- Full likelihood involves inversion of matrices - computationally difficult for large $n$
- Composite likelihood breaks up the information into smaller contributions making computations easier
- Can be specified marginally or conditionally
Methodology
Notation

- Let observations $Z$ observed at locations $s_i$ for $i = 1, ..., n$ be denoted as $Z(s_i)$ such that
  \[ Z \sim N(\mu, C(\sigma^2, \rho)) \]
- $\mu$ is mean, $\sigma^2$ is variability, $\rho$ is spatial dependence
- Process can be described by semivariogram which is function of covariance parameters as
  \[ \gamma(s_i, s_j) = \frac{1}{2} \text{var}(Z(s_i) - Z(s_j)) \]
- Special case of exponential semivariogram
  \[ \gamma(|s_i - s_j|; \phi) = c_0 + \sigma^2 \left( 1 - \exp^{-\frac{|s_i - s_j|}{\phi}} \right) \]
- Interest in $\mu, \sigma^2, \rho = \exp^{-1/\phi}$
Maximum likelihood & Restricted maximum likelihood

- **ML**
  
  \[
  l_{ML}(\phi, \mu; Z) = -\frac{1}{2} \ln |C(\phi)| - \frac{1}{2} (Z - \mu)^T C(\phi)^{-1} (Z - \mu)
  \]

- **REML** (Harville, D. 1974)
  
  \[
  l_{REML}(\phi, \mu; Z) = l_{ML}(\phi, \mu; Z) - \frac{1}{2} \ln |X' C(\phi)^{-1} X|
  \]

- Require full specification of the probabilistic model
- Involve inversion of matrices, which can become computationally prohibitive for large \( n \)
Composite likelihood

Pairwise differences ($\text{CL}_1$) (Curriero and Lele, 1999)

- Let $v_{ij} = Z(s_i) - Z(s_j)$ be the vector contrasts between pairs.

$$
\text{CL}(\phi; \mathbf{V}) = \prod_{i=1}^{n-1} \prod_{j>i}^{n} f(v_{ij}; \phi)
$$

- Assume that $v_{ij} \sim N(0, 2\gamma(d_{ij}; \phi))$, where $d_{ij} = |s_i - s_j|$ and $\gamma(\cdot)$ is the semivariogram function

$$
\text{cl}_D(\phi; \mathbf{Z}) = \sum_{i=1}^{n-1} \sum_{j>i}^{n} \left\{ \frac{(Z(s_i) - Z(s_j))^2}{4\gamma(d_{ij}; \phi)} + \frac{1}{2} \ln(\gamma(d_{ij}; \phi)) \right\}
$$

- Mean parameters treated as nuisance
Composite likelihood

Pairwise marginal (CL$_2$) (Le Cessie and Van Houwelingen, 1994)

- Let $Z_{ij} = (Z(s_i), Z(s_j))$ be the vector of all possible pairs between sites $i$ and $j$

$$CL(\phi, \mu; Z) = \prod_{i=1}^{n-1} \prod_{j > i}^{n} f(Z(s_i), Z(s_j); \phi, \mu_{ij})$$

- Log likelihood function

$$cl_{PM}(\phi, \mu; Z) = \sum_{i=1}^{n-1} \sum_{j > i}^{n} \ln \left( f(Z(s_i), Z(s_j); \phi, \mu_{ij}) \right)$$

$$= \sum_{i=1}^{n-1} \sum_{j > i}^{n} \ln \left( \frac{1}{\sqrt{(2\pi)^2 |C_{ij}|}} \exp \left( -\frac{1}{2} (Z_{ij} - \mu_{ij})^T C_{ij}^{-1} (Z_{ij} - \mu_{ij}) \right) \right)$$


Restricted composite likelihood (RECL)

- Let $Z_{ij} = (Z(s_i), Z(s_j))$ be the vector of all possible pairs between sites $i$ and $j$
- Loglikelihood

$$
\text{cl}_{\text{RECL}} = \sum_{i=1}^{n-1} \sum_{j>i}^{n} \left( w^* \ln f (Z(s_i), Z(s_j); \mu_{ij}, \phi) \right) - \frac{1}{2} \ln \left| \sum_{i=1}^{n-1} \sum_{j>i}^{n} X_{ij}' C_{ij}^{-1} X_{ij} \right|
$$

- $w^* = \frac{n'}{n(n-1)}$ derived from effective sample size $n'$ (Fortin and Dale, 2005)

$$
n' = \frac{n^2}{\sum_{i=1}^{n} \sum_{j=1}^{n} \rho(s_i, s_j)}
$$

- Correlation to use for effective sample size estimated
Weighting

- unweighted: $w_{ij} = 1$
- inverse distance: $w_{ij} = \frac{1}{d_{ij}}$
- neighbouring pairs: $w_{ij} = 1$ if neighbours, 0 otherwise
  - sharing of borders
  - distance between centroids
Standard error estimation

- Fisher information matrix

\[
(\hat{\theta} - \theta) \sim N\left(0, H(\theta)^{-1}\right)
\]

- Sandwich estimator using Godambe (1960) information matrix

\[
(\hat{\theta} - \theta) \sim N\left(0, H(\theta)^{-1}J(\theta)H(\theta)^{-1}\right)
\]

- \(J(\theta)\) dependent on the structure of the data - need pseudo-independent replications so that

\[
J(\theta) = \sum_{i=1}^{n} \sum_{j>i}^{n} \nabla cl(\hat{\theta}; y_i, y_j) \nabla cl(\hat{\theta}; y_i, y_j)^T.
\]
Standard error estimation

- Window subsampling
  - split region into (non)overlapping subregions $R_b$

  $$J_{WS} = \frac{1}{B} \sum_{b=1}^{B} |R_b| \nabla cl(\hat{\theta}; y \in R_b) \nabla cl(\hat{\theta}; y \in R_b)^T$$

- needs good mixing properties

- Monte Carlo simulations
  - generate new data $y^{(b)}$ using estimates $\hat{\theta}$

  $$J_{MC} = \frac{1}{B} \sum_{b=1}^{B} \nabla cl(\hat{\theta}; y^{(b)}) \nabla cl(\hat{\theta}; y^{(b)})^T$$

- assumes fitted model is correct
Simulation study
Simulation study setting

- Choose grid type $S_1$ ($n_1 = 64$, $n_2 = 225$, $n_3 = 225$), variance parameter $\sigma^2$, and spatial dependence level $\rho$ (weak, moderate, strong)

- **Step 1.** Simulate the data set $Z_1$ from an exponential correlation model with mean $\mu = \beta_0 + \beta_1X$.

- **Step 2.** Based on $Z_1$ estimate the model parameters using the different methods highlighted (ML, REML, $CL_1$, $CL_2$, $RECL_{true}$, $RECL_{est}$).

- **Step 3.** Repeat Steps 1 and 2 500 times for each grid type and dependence level combination specified.

True values: $\sigma^2 = 1$, $\beta_0 = 3$, $\beta_1 = -1$ where $X \sim U(0, 1)$
Results 8x8 regular grid

top panel - weak correlation, bottom panel moderate correlation

Variance

Correlation

Intercept

Covariate

1:ML, 2:REML, 3:CL₁, 4:CL₂, 5:RECL_true, 6:RECL_est
Results 15x15 increasing domain

top panel - weak correlation,     bottom panel moderate correlation

Variance

Correlation

Intercept

Covariate

1:ML, 2:REML, 3:CL\textsubscript{1}, 4:CL\textsubscript{2}, 5:RECL\textsubscript{true}, 6:RECL\textsubscript{est}
Results

- Weighting improves estimates – less bias
- $\hat{\rho}$ from differences best choice when true $\rho$ is not known
- Larger differences for strong correlations
- Variability reduces as $n$ increases – consistency
spatialRECL package
**spatialRECL package**

- Exponential correlation structure
- **Input:**
  - data in geodata format (shape polygon may also be included)
  - maximum distance
  - variance estimation method
- **Output:**
  - estimates of mean and variance parameters
  - subsampling regions used (where applicable)
  - neighbourhood and pairing information
- Still in development, will be made available on [http://repos.openanalytics.eu/](http://repos.openanalytics.eu/)

```r
fit <- fitRECL(formula = data ~ lon, geodata = wheat1, maxdist = 7, vartype = "subsampling", polygonData = wheat)
```
Application to data
Wheat yield (Mercer and Hall)

- 20 × 25 lattice of plots
- Harvested wheat in plots and weighed the yield
- Non-homogeneity in plots
- Covariate considered is longitude coordinate
- Neighbourhood defined by radius of 10
Results

<table>
<thead>
<tr>
<th></th>
<th>$\sigma^2$</th>
<th>$\rho$</th>
<th>Intercept</th>
<th>Covar</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>0.1925(NA)</td>
<td>0.6445(NA)</td>
<td>4.1544(0.0754)</td>
<td>-0.0064(0.002)</td>
</tr>
<tr>
<td>REML</td>
<td>0.1953(NA)</td>
<td>0.6518(NA)</td>
<td>4.1525(0.0774)</td>
<td>-0.0063(0.002)</td>
</tr>
<tr>
<td>CL$_2$</td>
<td>0.1884(0.0018)</td>
<td>0.6595(0.0142)</td>
<td>4.2119(0.0067)</td>
<td>-0.0081(0.0002)</td>
</tr>
<tr>
<td>RECL</td>
<td>0.2005(0.0303)</td>
<td>0.7179(0.1047)</td>
<td>4.2115(0.0457)</td>
<td>-0.0081(0.0016)</td>
</tr>
</tbody>
</table>

- likfit() function in geoR does not return standard errors for covariance parameters
TSH levels in Galicia

- 297 municipalities
- TSH levels in newborns measured at birth
- Covariate considered is mean birth weight in municipality
- Neighbourhood defined by radius of 0.5
## Results

<table>
<thead>
<tr>
<th></th>
<th>$\sigma^2$</th>
<th>$\rho$</th>
<th>Intercept</th>
<th>Covar</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>2.2083 (NA)</td>
<td>0 (NA)</td>
<td>4.0176 (1.3844)</td>
<td>-0.2545 (0.4202)</td>
</tr>
<tr>
<td>REML</td>
<td>2.2276 (NA)</td>
<td>0 (NA)</td>
<td>4.0145 (1.3878)</td>
<td>-0.2535 (0.4213)</td>
</tr>
<tr>
<td>CL$^2$</td>
<td>2.2038 (0.0245)</td>
<td>0 (0)</td>
<td>4.539 (0.1972)</td>
<td>-0.4274 (0.0599)</td>
</tr>
<tr>
<td>RECL</td>
<td>2.2253 (0.2644)</td>
<td>0 (0)</td>
<td>4.5368 (0.925)</td>
<td>-0.4268 (0.2661)</td>
</tr>
</tbody>
</table>
Conclusions

- Penalization seems important for pairwise marginal composite likelihood
- Less computation time than full likelihood
  - 10000 spatial points on grid
  - 2.93GHz, 24GB RAM computer
  - Full maximum likelihood - 5hrs
  - RECL on pairs within a neighbourhood distance - 15min
- New methods improve estimates of covariance parameters
- Important to use correct methods for standard error estimation
References


THANK YOU

kathy.mutambanengwe@openanalytics.eu