spmodel: Spatial Modeling in R – Supplementary Material

A Preprint

Michael Dumelle *
United States
Environmental Protection Agency
200 SW 35th St, Corvallis, OR, 97333
Dumelle.Michael@epa.gov

Matt Higham
Department of Math, Computer Science, and Statistics
St. Lawrence University
23 Romoda Drive, Canton, NY, 13617
mhigham@stlawu.edu

Jay M. Ver Hoef
National Oceanic and Atmospheric Administration
Alaska Fisheries Science Center
Marine Mammal Laboratory, Seattle, WA, 98115
jay.verhoef@noaa.gov

April 23, 2022

Abstract

Enter the text of your abstract here.

***K*eywords** Spatial covariance · Linear Model · Autoregressive model

1 Covariance Functions

2 Estimation

2.1 Likelihood-based Estimation

Minus twice a profiled Gaussian log-likelihood, denoted $-2l(\boldsymbol{\theta}|\mathbf{y})$ is given by

$$-2l(\boldsymbol{\theta}|\mathbf{y}) = \ln|\boldsymbol{\Sigma}| + (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) + n \ln 2\pi, \tag{1}$$

where $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^\intercal \boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^\intercal \boldsymbol{\Sigma}^{-1}\mathbf{y}$. Minimizing Equation 1 yields $\hat{\boldsymbol{\theta}}_{ml}$, the maximum likelihood estimates for $\boldsymbol{\theta}$. Then a closed for solution exists for $\hat{\boldsymbol{\beta}}_{ml}$, the maximum likelihood estimates for $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}}_{ml} = \tilde{\boldsymbol{\beta}}_{ml}$, where $\tilde{\boldsymbol{\beta}}_{ml}$ is $\tilde{\boldsymbol{\beta}}$ evaluated at $\hat{\boldsymbol{\theta}}_{ml}$. Unfortunately $\hat{\boldsymbol{\theta}}_{ml}$ can be badly biased for $\boldsymbol{\theta}$ (especially for small sample sizes), which impacts the estimation of $\boldsymbol{\beta}$ (Patterson and Thompson 1971). This bias occurs due to the simultaneous estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ To reduce this bias, restricted maximum likelihood estimation (REML) emerged (Patterson and Thompson 1971; Harville 1977; Wolfinger, Tobias, and Sall 1994). It can be shown that integrating $\boldsymbol{\beta}$ out of a Gaussian likelihood yields the restricted Gaussian likelihood used in REML estimation. Minus twice a restricted Gaussian log-likelihood, denoted $-2l_R(\boldsymbol{\theta}|\mathbf{y})$ is given by

$$-2l_R(\boldsymbol{\theta}|\mathbf{y}) = -2l(\boldsymbol{\theta}|\mathbf{y}) + \ln|\mathbf{X}^\intercal \boldsymbol{\Sigma}^{-1}\mathbf{X}| - p \ln 2\pi, \tag{2}$$

where p equals the dimension of $\boldsymbol{\beta}$. Minimizing Equation 2 yields $\hat{\boldsymbol{\theta}}_{reml}$, the restricted maximum likelihood estimates for $\boldsymbol{\theta}$. Then a closed for solution exists for $\hat{\boldsymbol{\beta}}_{reml}$, the restricted maximum likelihood estimates for $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}}_{reml} = \tilde{\boldsymbol{\beta}}_{reml}$, where $\tilde{\boldsymbol{\beta}}_{reml}$ is $\tilde{\boldsymbol{\beta}}$ evaluated at $\hat{\boldsymbol{\theta}}_{reml}$.

*Corresponding Author

Generally the overall variance, σ^2, can be profiled out of Equation 1 and Equation 2. This reduces the number of parameters requiring optimization by one, which can dramatically reduce estimation time. For example, profiling σ^2 out of Equation 1 yields

$$-2l^*(\boldsymbol{\theta}^*|\mathbf{y}) = \ln|\boldsymbol{\Sigma}^*| + n\ln[(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})] + n + n\ln 2\pi/n. \tag{3}$$

After finding $\hat{\boldsymbol{\theta}}^*_{ml}$ a closed form solution for $\hat{\sigma}^2_{ml}$ exists: $\hat{\sigma}^2_{ml} = [(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})]/n$. Then $\hat{\boldsymbol{\theta}}^*_{ml}$ is combined with $\hat{\sigma}^2_{ml}$ to yield $\hat{\boldsymbol{\theta}}_{ml}$ and subsequently $\hat{\boldsymbol{\beta}}_{ml}$. A similar result holds for REML estimation. Profiling σ^2 out of Equation 2 yields

$$-2l^*_R(\boldsymbol{\theta}^*|\mathbf{y}) = \ln|\boldsymbol{\Sigma}^*| + (n-p)\ln[(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})] + (n-p) + (n-p)\ln 2\pi/(n-p). \tag{4}$$

After finding $\hat{\boldsymbol{\theta}}^*_{reml}$ a closed form solution for $\hat{\sigma}^2_{reml}$ exists: $\hat{\sigma}^2_{reml} = [(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})]/(n-p)$. Then $\hat{\boldsymbol{\theta}}^*_{reml}$ is combined with $\hat{\sigma}^2_{reml}$ to yield $\hat{\boldsymbol{\theta}}_{reml}$ and subsequently $\hat{\boldsymbol{\beta}}_{reml}$.

2.2 Semivariogram-based Estimation

An alternative approach to likelihood-based estimation is semivariogram-based estimation. The semivariogram of a constant-mean process $\mathbf{y}$ is the expectation of the squared half-difference between two observations h distance units apart. More formally, the semivariogram is denoted $\gamma(h)$ and defined as

$$\gamma(h) = \mathrm{E}(y_i - y_j)^2/2, \tag{5}$$

where $||y_i - y_j||_2 = h$ (the Euclidean distance). When the process $\mathbf{y}$ is second-order stationary, the semivariogram and covariance function are intimately connected: $\gamma(h) = \mathrm{Cov}(0) - \mathrm{Cov}(h)$, where $\mathrm{Cov}(0)$ is the covariance function evaluated at 0 (which is the overall variance, σ^2) and $\mathrm{Cov}(h)$is the covariance function evaluated at h.

2.2.1 Weighted Least Squares

The empirical semivariogram is a moment-based estimate of the semivariogram denoted by $\hat{\gamma}(h)$ and defined as

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|}\sum_{N(h)}(y_i - y_j)^2, \tag{6}$$

where $N(h)$ is the set of observations in $\mathbf{y}$ that are h units apart (distance classes) and $|N(h)|$ is the cardinality of $N(h)$ (Cressie 1993). Often the set $N(h)$ contains observations that are $h \pm \alpha$ apart – this approach is known as "binning" the empirical semivariogram. Equation (6) is viewed as the average squared half-distance between two observations in $\mathbf{y}$. Cressie (1985) proposed estimating $\boldsymbol{\theta}$ by minimizing an objective function that involves γh and $\hat{\gamma}(h)$ and is based on a weighted least squares criterion. This criterion is defined as

$$\sum_i w_i[\hat{\gamma}(h)_i - \gamma(h)_i]^2, \tag{7}$$

where w_i, $\hat{\gamma}(h)_i$, and $\gamma(h)_i$ are the weights, empirical semivariogram, and semivariogram for the ith distance class. Cressie (1985) recommended setting $w_i = |N(h)|/\gamma(h)^2_i$, which gives more weights to distance classes with more observations ($|N(h)|$) and semivariances at shorter distances ($1/\gamma(h)^2_i$). The default in spmodel is to use these w_i – the type of w_i is changed via the `weights` argument to `splm()`. Table 2.2.1 contains all w_i available in spmodel.

Recall that the semivariogram is defined for a constant-mean process. Typically in linear models, $\mathbf{y}$ does not have a constant mean. So the empirical semivariogram and $\hat{\boldsymbol{\theta}}_{wls}$ are actually constructed using the residuals from an ordinary least squares regression of $\mathbf{y}$ on $\mathbf{X}$ – these residuals are assumed to have mean zero.

2.2.2 Composite Likelihood

(Curriero and Lele 1999)

w_i Name	w_i Form	`weight =`
Cressie	$\|N(h)\|/\gamma(h)_i^2$	`"cressie"`
Cressie (Denominator) Root	$\|N(h)\|/\gamma(h)_i$	`"cressie-droot"`
Cressie No Pairs	$1/\gamma(h)_i^2$	`"cressie-nopairs"`
Cressie (Denominator) Root No Pairs	$1/\gamma(h)_i$	`"cressie-droot-nopairs"`
Pairs	$\|N(h)\|$	`"pairs"`
Pairs Inverse Distance	$\|N(h)\|/h^2$	`"pairs-invd"`
Pairs Inverse (Root) Distance	$\|N(h)\|/h$	`"pairs-invsd"`
Ordinary Least Squares	1	`ols`

Table 1: spmodel table weights

3 Hypothesis Testing

3.1 The General Linear Hypothesis Test

3.2 Contrasts

4 Random Effects

4.1 BLUPs

References

Cressie, Noel. 1985. "Fitting Variogram Models by Weighted Least Squares." *Journal of the International Association for Mathematical Geology* 17 (5): 563–86.

———. 1993. *Statistics for Spatial Data.* John Wiley & Sons.

Curriero, Frank C, and Subhash Lele. 1999. "A Composite Likelihood Approach to Semivariogram Estimation." *Journal of Agricultural, Biological, and Environmental Statistics*, 9–28.

Harville, David A. 1977. "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems." *Journal of the American Statistical Association* 72 (358): 320–38.

Patterson, H Desmond, and Robin Thompson. 1971. "Recovery of Inter-Block Information When Block Sizes Are Unequal." *Biometrika* 58 (3): 545–54.

Wolfinger, Russ, Randy Tobias, and John Sall. 1994. "Computing Gaussian Likelihoods and Their Derivatives for General Linear Mixed Models." *SIAM Journal on Scientific Computing* 15 (6): 1294–1310.