SPMODEL: SPATIAL MODELING IN **R** – COMPUTATINAL DETAILS VIGNETTE

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

May 31, 2022

Abstract

Fit, summarize, and predict for a variety of spatial models. Parameters of spatial linear models and spatial autoregressive models are estimated using a variety of methods. Additional modeling features include anisotropy, random effects, partition factors, big data approaches, and more. Model-fit statistics are used to summarize, visualize, and compare models. Predictions at unobserved locations are easily obtainable.

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

1 Introduction

This vignette covers technical details regarding the functions in spmodel that perform computations.

A note on square roots and Cholesky solves

2 `splm()` and `spautor()`

2.1 'splm()

3 Additional Functions'

3.0.1 Covariance Functions

3.0.2 Estimation

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

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

---

[*]Corresponding Author

where $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{y}$. Minimizing Equation 3.0.2 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

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

where p equals the dimension of $\boldsymbol{\beta}$. Minimizing Equation 1 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 $\hat{\boldsymbol{\theta}}_{reml}$.

Generally the overall variance, σ^2, can be profiled out of Equation 3.0.2 and Equation 1. This reduces the number of parameters requiring optimization by one, which can dramatically reduce estimation time. For example, profiling σ^2 out of Equation 3.0.2 yields

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

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 1 yields

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

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 $\hat{\boldsymbol{\beta}}_{reml}$.

Both ML and REML estimation rely on the an $n \times n$ covariance matrix inverse. Inverting an $n \times n$ matrix is an enormous computational demand that scales cubically with the sample size. For this reason, ML and REML have historically been unfeasible to implement in their standard form with data larger than a few thousand observations. This motivates the use for the big data approaches outlined in Section (INSERT SECTION).

It is worth noting that the inverses themselves are not strictly needed for estimation (or prediction), though at least their square root is needed. In spmodel, calculating this square root requires a Cholesky decomposition, which still scales cubically with the sample size. Computing the Cholesky decomposition, however, is far more computationally efficient than computing the inverse. To see why only the Cholesky decomposition is needed, recall that the Cholesky decomposition of the covariance matrix $\boldsymbol{\Sigma}$ is $\mathbf{C}\mathbf{C}^\intercal$, where $\mathbf{C}$ is a lower triangular matrix (so $\mathbf{C}\mathbf{C}^\intercal = \boldsymbol{\Sigma}$). In the ML and REML likelihoods, $\boldsymbol{\Sigma}^{-1}$ is not needed on its own, only $\mathbf{X}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{X}$ and $\mathbf{X}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{y}$ are needed. We can rewrite the $\mathbf{X}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{X}$ as $\mathbf{X}^\intercal(\mathbf{C}^\intercal)^{-1}\mathbf{C}^{-1}\mathbf{X} = (\mathbf{C}^{-1}\mathbf{X})^\intercal\mathbf{C}^{-1}\mathbf{X}$. Then $\mathbf{C}^{-1}\mathbf{X}$ is efficiently solved by noticing that $\mathbf{C}^{-1}\mathbf{X} = \mathbf{A}$ for some matrix $\mathbf{A}$ implies $\mathbf{X} = \mathbf{C}\mathbf{A}$. This system can be efficiently solved for $\mathbf{A}$ using linear forward solves (forward substitution). Then $\mathbf{X}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{X} = \mathbf{A}^\intercal\mathbf{A}$. A similar approach is used to solve $\mathbf{X}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{y}$. Still, using Cholesky decomositions is unfeasible for sample sizes larger than a few thousand observations.

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

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

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. Both semivariogram approaches described next are more computationally efficient than ML or REML because their major computational burden (calculations based on pairs) scale the squared sample size (i.e., not the cubed sample size).

2

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

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

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

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. Typically, only certain h considered when constructing Equation 2 – a commonly used cutoff is to ignore h larger than half the maximum distance in the domain. One criticism of the empirical semivariogram is that distance bins and cutoffs tend to be artitrarily chosen (i.e., not chosen according to some statistical criteria).

Equation (2) 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

$$\sum_i w_i [\hat{\gamma}(h)_i - \gamma(h)_i]^2, \tag{3}$$

where w_i, $\hat{\gamma}(h)_i$, and $\gamma(h)_i$ are the weights, empirical semivariogram, and semivariogram for the ith distance class. Minimizing Equation (3) yields $\hat{\boldsymbol{\theta}}_{wls}$, the semivariogram weighted least squares estimates of $\boldsymbol{\theta}$. After estimating $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ estimates are constructed using (empirical) generalized least squares: $\hat{\boldsymbol{\beta}}_{wls} = (\mathbf{X}^\intercal \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\intercal \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{y}$, where $\hat{\boldsymbol{\Sigma}}^{-1}$ is $\boldsymbol{\Sigma}$ evaluated at $\hat{\boldsymbol{\theta}}_{wls}$.

Cressie (1985) recommends setting the w_i in Equation (3) as $w_i = |N(h)|/\gamma(h)_i^2$, which gives more weights to distance classes with more observations ($|N(h)|$) and semivariances at shorter distances ($1/\gamma(h)_i^2$). The default in spmodel is to use these w_i – the type of w_i is changed via the `weights` argument to `splm()`. Table 3.0.2 contains all w_i available in spmodel.

Additionally, the number of $N(h)$ classes and maximum distance for h are specified by passing the `bins` and `cutoff` arguments to `splm()` (these arguments are passed via `...` to `esv()`). The default value for `bins` is 15 and the default value for the maximum h is half the maximum distance of the spatial domain's bounding box.

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

Composite Likelihood The composite likelihood approach involves constructing a likelihood based on conditional or marginal events for which log-likelihoods are available and then adding together these individual components. Composite likelihoods are attractive because they behave very similar to likelihoods but are easier to handle, both from a theoretical and a computational perspective. Curriero and Lele (1999) derive a particular composite likelihood for estimating semivariogram parameters. The negative log of this composite likelihood, denoted CL(h), is given by

$$\mathrm{CL}(h) = \sum_{i=1}^{n-1} \sum_{j>i} \left(\frac{(y_i - y_j)^2}{2\gamma(h)} + \ln(\gamma(h)) \right)$$
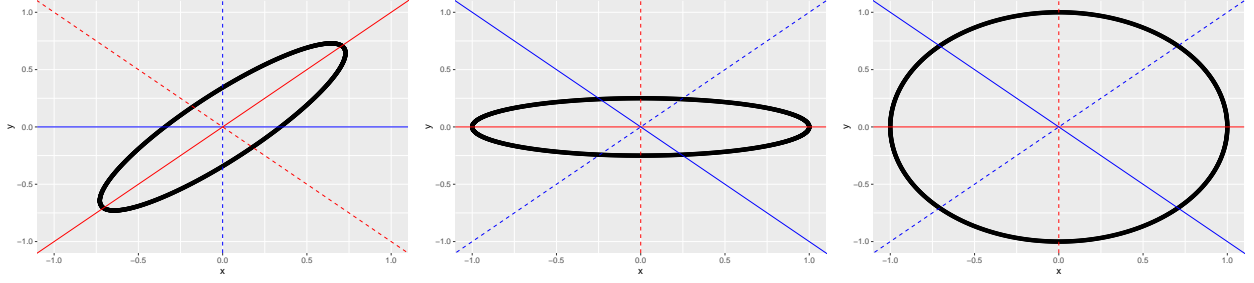
Figure 1: In the first figure, the ellipse of an anisotropy covariance is shown. In the second figure, the ellipse is rotated so the major axis is the transformed x-axis and the minor axis is the transformed y-axis. In the third figure, the ellipse is scaled so the ellipse becomes a circle. Then covariance estimation can be performed (using the transformed coordinates that compose the third figure).

where $\gamma(h)$ is the semivariogram (that depends on parameter vector $\boldsymbol{\theta}$). Minimizing Equation 3.0.2 yields $\hat{\boldsymbol{\theta}}_{cl}$, the semivariogram composite likelihood estimates of $\boldsymbol{\theta}$. After estimating $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ estimates are constructed using (empirical) generalized least squares: $\hat{\boldsymbol{\beta}}_{cl} = (\mathbf{X}^{\intercal}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$, where $\hat{\boldsymbol{\Sigma}}^{-1}$ is $\boldsymbol{\Sigma}$ evaluated at $\hat{\boldsymbol{\theta}}_{cl}$.

An advantage of the composite likelihood approach to semivariogram estimation is that it does not require arbitrarily specifying empircial semivariogram bins and cutoffs. It does tend to be more computationally demanding than the weighted least squares, however, as the composite likelihood is constructed from $\binom{n}{2}$ pairs for a sample size n, whereas the weighted least squares approach only requires calculating $\binom{|N(h)|}{2}$ pairs for each distance bin $N(h)$. As with the weighted least squares approach, Equation 3.0.2 requires constant-mean process, so typically the residuals from an ordinary least squares regression of $\mathbf{y}$ on $\mathbf{X}$ are used to estimate $\boldsymbol{\theta}$.

3.0.3 Optimization

3.0.4 Hypothesis Testing

3.0.5 (Geometric) Anisotropy

A spatial process is geometrically isotropic if its covariance decays equally in all directions. A spatial process is geometrically anisotropic if its covariance does not decay equally in all directions (i.e., is not isotropic). Let h represent distance, α represent an angle in $[0, \pi]$, and τ represent a scaling factor in $[0, 1]$. Geometric anisotropy is accommodate through a specific rotation and scaling of the x-coordinate and y-coordinate. Then a new distance, $h*$, corresponds to an isotropic process. The scaling works by rotating the major axis of the covariance clockwise onto a transformed x-axis and scaling the minor axis of the covariance along the transformed y-axis. Geometrically, this transformation and re-scaling is achieved via

$$\begin{bmatrix}\mathbf{x}^* \\ \mathbf{y}^*\end{bmatrix} = \begin{bmatrix}1 & 0 \\ 0 & 1/\tau\end{bmatrix}\begin{bmatrix}\cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha)\end{bmatrix}\begin{bmatrix}\mathbf{x} \\ \mathbf{y}\end{bmatrix}$$

Then covariance parameters are estimated using the transformed coordinates to compute distances. An anisotropy transformation is visualized in Figure 1.

In spmodel, anisotropy parameters (α and τ) can be estimated using ML or REML by setting the `anisotropy` argument to `TRUE` or by specifying one of the anisotropy parameters as unknown or known and different from zero in `spcov_initial`. Fixed anisotropy parameters can be provided to SV-WLS and SV-CL and incorporated into estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ when specified in `spcov_initial`.

Estimating anisotropy can be challenging because of local minima in the likelihood. First note that estimation need only occur for α on $(0, \pi)$ radians due to symmetric of the covariance ellipse at rotations α and $\alpha + \pi$. Second note that estimation need only occur for τ on $(0, 1)$ because we have defined τ as the scaling factor for the length of the minor axis relative to the major axis. To address the local minima problem, each optimization iteration actually involves two likelihood evaluations, one for $\alpha = \alpha_0$ and another for $pi - \alpha_0$, so that the likelihood is evaluated for α parameters the quadrants on either side of the minor (transformed y) axis.

3.0.6 The Local list

3.1 `spautor()`

4 Additional Functions

4.1 `AIC()` and `AICc()`

The `AIC()` and `AICc()` functions in spmodel follow Hoeting et al. (2006), who define spatial AIC and AICc as

$$\text{AIC} = -2\ln(\hat{L}) + 2(npar + 1)$$
$$\text{AICc} = -2\ln(\hat{L}) + 2n(npar + 1)/(n - npar - 2),$$

where $\ln(\hat{L})$ is the log likelihood evaluated at its maximum given the estimated parameters, $npar$ is the number of estimated parameters, and n is the sample size. AIC and AICc are only defined for restricted maximum likelihood and maximum likelihood estimation, as they are the only estimation methods in spmodel that maximize a likelihood. For restricted maximum likelihood, $npar$ is the number of estimated covariance parameters. For maximum likelihood, $npar$ is the number of estimated covariance parameters plus the number of estimated fixed effects. The discrepancy arises because restricted maximum likelihood integrates the fixed effects out of the likelihood, thus the likelihood does not depend on the fixed effects (though it does still depend on the predictor variables).

4.2 `anova()`

4.2.1 The General Linear Hypothesis Test

Test statistics from `anova()` are formed using the general linear hypothesis test. Let $\mathbf{L}$ be an $l \times p$ contrast matrix and l_0 be an $l \times 1$ vector. The null hypothesis is that $\mathbf{L}\hat{\boldsymbol{\beta}} = l_0$ and the alternative hypothesis is that $\mathbf{L}\hat{\boldsymbol{\beta}} \neq l_0$. Usually, l_0 is the zero vector (and in spmodel, this is assumed). The test statistic, denoted X^2, is given by

$$\tilde{F} = [(\mathbf{L}\hat{\boldsymbol{\beta}} - l_0)^\intercal (\mathbf{L}(\mathbf{X}^\intercal \hat{\boldsymbol{\Sigma}} \mathbf{X})^{-1} \mathbf{L}^\intercal)^{-1} (\mathbf{L}\hat{\boldsymbol{\beta}} - l_0)]/rank(\mathbf{L})$$

By default, $\mathbf{L}$ is chosen such that each variable in the data used to fit the model is tested. If this deafult is not desired, the `Terms` and `L` arguments can be used to pass user-defined $\mathbf{L}$ matrices to `anova()`.

It is notoriously difficult to determine appropriate p-values for linear mixed models based on the general linear hypothesis test. lme4, for example, does not report p-values by default. There are three reasons we focus on next that explain why obtaining p-values is so challenging.

- The first (and often most important) challenge is that when estimating $\boldsymbol{\theta}$, it is usually not clear what the null distribution of $\tilde{F}$. In certain cases such as ordinary least squares regression or certain experimental designs (e.g., blocked design, split plot design, etc.), $\tilde{F}$ is F-distributed with known numerator and denominator degrees of freedom. But outside of these well-studied cases, no general results exist.
- The second challenge is that the standard error of $\tilde{F}$ does not account for the uncertainty in $\hat{\boldsymbol{\theta}}$. For some approaches to addressing this problem, see Kackar and Harville (1984), Prasad and Rao (1990), Harville and Jeske (1992), and Kenward and Roger (1997).
- The third challenge is in determining denominator degrees of freedom. Again, in certain cases, these are known – but this is not true in general. For some approaches to addressing this problem, see Satterthwaite (1946), Schluchter and Elashoff (1990), Hrong-Tai Fai and Cornelius (1996), Kenward and Roger (1997), Littell et al. (2006), Pinheiro and Bates (2006), and Kenward and Roger (2009).

For these reasons, spmodel assumes a large-sample, Chi-squared approximation when calculating p-values. This approach addresses the three points above by assuming that with a large enough sample size:

- The numerator of $\tilde{F}$ tends to be asymptotically Chi-squared (under certain conditions) with $rank(\mathbf{L})$ degrees of freedom.
- The uncertainty from estimating $\hat{\boldsymbol{\theta}}$ is small enough to be safely ignored.

Because the approximation is asymptotic, degree of freedom adjustments can be ignored (it is also worth noting that an F distribution with infinite denominator degrees of freedom is a scaled (by $rankL$) Chi-squared

distribution). A takeaway here is that this asymptotic approximation implies these p-values are likely unreliable with small samples.

A second approach to determining p-values is a likelihood ratio test for nested models. Let $l(\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\beta}}_0|\mathbf{y})$ be the log-likelihood from some reduced model and $l(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\beta}}_1|\mathbf{y})$ be the log-likelihood from some full model. When the reduced model is nested in the full model (i.e., the reduced model can be obtained by fixing some parameters of the full model), $[-2l(\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\beta}}_0|\mathbf{y})] - [-2l(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\beta}}_1|\mathbf{y})]$ is asymptotically Chi-squared with degrees of freedom equal to the difference in estimated parameters between the full and reduced model.

For REML estimation, likelihood ratio tests can only be used to compare nested models whose fixed effect structure does not change. This is because the REML likelihood (1) depends on the fixed effects through $\ln|\mathbf{X}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{X}|$. To use likelihood ratio tests for assessing the importance of fixed effects, parameters must be estimated using ML. When using likelihood ratio tests to assess the importance of parameters on the boundary of a parameter space (e.g., a variance parameter being zero), p-values tend to be too large (Self and Liang 1987; Stram and Lee 1994; Goldman, Whelan, and Simon 2000; Pinheiro and Bates 2006).

4.3 `coef()`

`coef()` returns relevant coefficients based on the `type` argument. When `type = "fixed"` (the default), `coef()` returns $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}^\intercal\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$, where $\hat{\boldsymbol{\Sigma}}^{-1}$ is the inverse covariance matrix evaluated at the estimated covariance parameters. If the estimation method is restricted maximum likelihood or maximum likelihood, this estimator is known as the restricted maximum likelihood or maximum likelihood estimator of $\boldsymbol{\beta}$. If the estimation method is semivariogram weighted least squares or semivariogram composite likelihood, this estimator is known as the empirical generalized least squares estimator of $\boldsymbol{\beta}$. When `type = "spcov"`, the estimated spatial covariance parameters are returned (available for all estimation methods). When `type = "randcov"`, the estimated random effect variance parameters are returned (available for restricted maximum likelihood and maximum likelihood estimation methods).

4.4 `confint()`

`confint()` returns confidence intervals for estimated parameters. Currently, `confint()` only returns confidence intervals for the fixed effects. The confidence interval for β_i is $\beta_i \pm z^*\sqrt{(\mathbf{X}^\intercal\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})_{ii}^{-1}}$, where $\Phi(z^*) = 1 - \alpha/2$, $\Phi(\cdot)$ is the standard normal (Gaussian) cumulative distribution function, $\alpha = 1-$ `level`, and $(\mathbf{X}^\intercal\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})_{ii}^{-1}$ is the ith diagonal element in $(\mathbf{X}^\intercal\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}$. The default for `level` is 0.95, which corresponds to a z^* of approximately 1.96.

4.5 `cooks.distance()`

Cook's distance measures the influence of an observation. For independent observations, the Cook's distance of a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is $\hat{\boldsymbol{\epsilon}}(\mathbf{H}_{ii}/[p(1 - \mathbf{H}_{ii})])$, where p is the number of fixed effects and $\mathbf{H}_{ii}$ is the hat matrix equal to $\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$. For dependent observations, the Cook's distance can be viewed as the standard Cook's distance of "whitened" observations corresponding to the linear model $\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$, where $\mathbf{y}^* = \boldsymbol{\Sigma}^{-1/2}\mathbf{y}$, $\mathbf{X}^* = \boldsymbol{\Sigma}^{-1/2}\mathbf{y}$, and $\boldsymbol{\epsilon}^* = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon}$. So the whitened Cook's distance is $(\hat{\boldsymbol{\epsilon}}_i^*(\mathbf{H}_{ii}^*/[p(1 - \mathbf{H}_{ii}^*)])$, where $\mathbf{H}_{ii}^*$ is the ith diagonal entry of the hat matrix $\mathbf{H}^* = \mathbf{X}^*[(\mathbf{X}^*)^\intercal\mathbf{X})]^{-1}(\mathbf{X}^*)^\intercal$ and $\hat{\boldsymbol{\Sigma}}$ is the covariance matrix evaluated at the estimated covariance parameters. Suggested cutoffs for identifying influential observations are $4/n$, where n is the sample size, or 1.

4.6 `deviance()`

The deviance of a fitted model is $2\ln(\hat{\mathbf{L}}_s) - 2\ln(\hat{\mathbf{L}}_f)$, where $2\ln(\hat{L}_s)$ is twice the log-likelihood of a "saturated" model that fits every observation perfectly and $2\ln(\hat{L}_f)$ is twice the log-likelihood of the fitted model. For normal (Gaussian) data, the deviance of a fitted model is $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\intercal\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$.

4.7 esv()

4.8 fitted()

Fitted values for the response, spatial random errors, and random effects are obtained in spmodel using `fitted()`. The fitted values for the response, denoted $\hat{\mathbf{y}}$, are given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$$

They are the estimated mean response given the set of covariates for each observation.

Fitted values for spatial random errors and random effects are linked with best linear unbiased predictors from mixed model theory. Consider the standard random effects parameterization

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon},$$

where $\mathbf{Z}$ denotes the random effects design matrix, $\mathbf{u}$ denotes the random effects, and $\boldsymbol{\epsilon}$ denotes independent random error. Henderson (1975) states that the best linear unbiased predictor (BLUP) of a single random effect $\mathbf{u}$, denoted $\hat{\mathbf{u}}$, is given by

$$\hat{\mathbf{u}} = \sigma_u^2 \mathbf{Z}\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

Searle, Casella, and McCulloch (2009) generalizes this idea by showing that for a random variable $\boldsymbol{\alpha}$, the best linear unbiased predictor (based on the response, $\mathbf{y}$) of α, denoted $\hat{\boldsymbol{\alpha}}$ is given by

$$\hat{\boldsymbol{\alpha}} = \mathrm{E}(\boldsymbol{\alpha}) + \Sigma_\alpha \Sigma^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{4}$$

where $\Sigma_\alpha = \mathrm{Cov}(\boldsymbol{\alpha}, \mathbf{y})$. Evaluating Equation (4) at the plug-in (empirical) estimates of the covariance parameters yields the empirical best linear unbiased predictor (EBLUP) of $\boldsymbol{\alpha}$.

Building from this idea and putting plug-in (empirical) estimates of the covariance parameters into Equation (4) yields predictors for each random error, which we call fitted values henceforth. For example, the fitted value corresponding to the spatial dependent random error, denoted $\hat{\boldsymbol{\tau}}$, is given by

$$\hat{\boldsymbol{\tau}} = \boldsymbol{\Sigma}_{de}\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

where $\boldsymbol{\Sigma}_{de} = \mathrm{Cov}(\boldsymbol{\tau}, \mathbf{y})$. To obtain these fitted values for the spatial random errors, run `fitted(object, type = "spcov")`. To obtain these fitted values for the random effects, run `fitted(object, type = "randcov")`.

(Talk about assumption of independent among the random effects yields the EBLUPs for us?)

4.9 hatvalues()

Hat values measure the leverage of an observation. For independent observations, the hat matrix of a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is $\mathbf{H} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$. The leverage of the ith observation is the ith diagonal element of $\mathbf{H}$. For dependent observations, the Hat matrix can be viewed as the standard Hat matrix of "whitened" observations corresponding to the linear model $\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$, where $\mathbf{y}^* = \boldsymbol{\Sigma}^{-1/2}\mathbf{y}$, $\mathbf{X}^* = \boldsymbol{\Sigma}^{-1/2}\mathbf{y}$, and $\boldsymbol{\epsilon}^* = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon}$. So the whitened Hat matrix is $\mathbf{H}^* = \mathbf{X}^*[(\mathbf{X}^*)^\intercal\mathbf{X}]^{-1}(\mathbf{X}^*)^\intercal$ and $\hat{\boldsymbol{\Sigma}}$ is the covariance matrix evaluated at the estimated covariance parameters. Suggested cutoffs for identifying observations with high leverage are $2p/n$ or $3p/n$, where p is the number of fixed effects and n is the sample size.

4.10 logLik()

`logLik()` returns the value of the log-likelihood for fitted models estimated using restricted maximum likelihood or maximum likelihood.

4.11 loocv()

4.12 plot()

4.13 predict()

Predictions are often the primary goal of a data analysis. spmodel performs best linear unbiased prediction, which is equivalent to Kriging, using the `predict()` function. Let $\mathbf{y}_o$ be a vector of observed variables and $\mathbf{y}_u$ be a vector of unobserved variables. Let $\mathbf{X}_o$ and $\mathbf{X}_u$ be the fixed effect design matrices of $\mathbf{y}_o$ and $\mathbf{y}_u$,

respectively. Let $\mathbf{\Sigma}_o$ and $\mathbf{\Sigma}_u$ be the covariance matrices of $\mathbf{y}_o$ and $\mathbf{y}_u$, respectively. The (empirical) best linear unbiased predictor (BLUP; Kriging predictor) of $\mathbf{y}_u$, denoted $\dot{\mathbf{y}}_{\mathbf{u}}$ is given by

$$\dot{\mathbf{y}}_{\mathbf{u}} = \mathbf{X}_u\hat{\boldsymbol{\beta}} + \hat{\mathbf{c}}_{uo}\hat{\mathbf{\Sigma}}_o^{-1}(\mathbf{y}_o - \mathbf{X}_o\hat{\boldsymbol{\beta}}),$$

where $\hat{\mathbf{c}}_{uo} = \hat{\mathrm{Cov}}(\mathbf{y}_u, \mathbf{y}_o)$ and $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\hat{\mathbf{\Sigma}}^{-1}\mathbf{X})^{-1}$. One can show that $\hat{\mathrm{Cov}}(\dot{\mathbf{y}})$ is

$$\mathrm{Cov}(\dot{\mathbf{y}}) = \hat{\mathbf{\Sigma}}_u - \hat{\mathbf{c}}_{uo}\hat{\mathbf{\Sigma}}_o^{-1}\hat{\mathbf{c}}_{uo}^\mathsf{T} + \mathbf{H}(\mathbf{X}^\mathsf{T}\hat{\mathbf{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{H}^\mathsf{T},$$

where $\mathbf{H} = (\mathbf{X}_u - \hat{\mathbf{c}}_{uo}\hat{\mathbf{\Sigma}}_o^{-1}\mathbf{X}_o)$. When `predict(..., interval = "none")`, predictions are obtained using Equation 4.13. If `se.fit = TRUE`, standard errors are the square root of the diagonal of Equation 4.13. When `predict(..., interval = "prediction")`, predictions are obtained using Equation 4.13 and standard errors used for the lower and upper bound of the prediction intervals are obtained via the square root of the diagonal of Equation 4.13 (these are also what are returned if `se.fit = TRUE`). When `predict(..., interval = "confidence")`, mean estimates are obtained using $\mathbf{X}_u\hat{\boldsymbol{\beta}}$ and standard errors used for the lower and upper bound of the confidence intervals are obtained via the square root of the diagonal of $(\mathbf{X}^\mathsf{T}\hat{\mathbf{\Sigma}}^{-1}\mathbf{X})^{-1}$ (these are also what are returned if `se.fit = TRUE`).

For autoregressive models, an extra step is required to obtain $\hat{\Sigma}_o^{-1}$ and $\hat{\Sigma}_u^{-1}$, as they depend on one another through the locations of $\mathbf{y}_o$ and $\mathbf{y}_u$. Let $\hat{\Sigma}^{-1}$ be the inverse covariance matrix of $\mathbf{y}_o$ and $\mathbf{y}_u$ (which we directly specify for autoregressive models). If the data are small enough, then $\hat{\Sigma}^{-1}$ can be inverted, yielding $\hat{\Sigma}$. Then appropriate subsets can then used to obtain the components needed for prediction. A faster way to obtain relevant components required for prediction does exist and is worth mentioning – $\hat{\Sigma}^{-1}$ can be represented blockwise as

$$\hat{\Sigma}^{-1} = \begin{bmatrix} \hat{\mathbf{\Sigma}}_{o,o}^* & \hat{\mathbf{\Sigma}}_{o,u}^* \\ \hat{\mathbf{\Sigma}}_{u,o}^* & \hat{\mathbf{\Sigma}}_{u,u}^* \end{bmatrix},$$

where the dimensions of the blocks are indexed by the matrix subscripts (rows then columns). All of the terms required for prediction can be obtained from this block representation. Wolf (1978) shows that

$$\hat{\Sigma}_o^{-1} = \hat{\mathbf{\Sigma}}_{o,o}^* - \hat{\mathbf{\Sigma}}_{o,u}^*(\hat{\mathbf{\Sigma}}_{u,u}^*)^{-1}\hat{\mathbf{\Sigma}}_{u,o}^*$$
$$\hat{\Sigma}_u = (\hat{\mathbf{\Sigma}}_{u,u}^* - \hat{\mathbf{\Sigma}}_{u,o}^*(\hat{\mathbf{\Sigma}}_{o,o}^*)^{-1}\hat{\mathbf{\Sigma}}_{o,u}^*)^{-1}$$
$$\mathrm{Cov}(\mathbf{y}_o, \mathbf{y}_u) = -\hat{\Sigma}_o\hat{\mathbf{\Sigma}}_{o,u}^*(\hat{\mathbf{\Sigma}}_{u,u}^*)^{-1}$$

A similar result exists for the log determinant (used in ML and REML estimation).

4.14 `pseudoR2()`

`pseudoR2()` is the pseudo r-squared equaling $1 - \mathbf{D}_f/\mathbf{D}_n$, where $\mathbf{D}_f$ is the deviance of the fitted model and $\mathbf{D}_n$ is the deviance of an intercept-only model (with the same estimated covariance parameters as for the fitted model). If the observations are independent and normal (Gaussian), this pseudo r-squared equals the classical r-squared $1 - SSE/SST$ for $SSE = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\mathsf{T}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ and $SST = (\mathbf{y} - \hat{\mu})^\mathsf{T}(\mathbf{y} - \mathbf{X}\hat{\mu})$ for intercept $\hat{\mu}$.

4.15 `residuals()`

`residuals()` returns various fitted-model residuals. When `type = "raw"`, the raw residuals are returned, $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. When `type = "normalized"`, the normalized residuals are returned, $\hat{\mathbf{\Sigma}}^{-1/2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$. When `type = "standardized"`, the standardized residuals are returned, $\hat{\mathbf{\Sigma}}^{-1/2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/\sqrt{(1 - diag(\mathbf{H}^*))}$, where $diag(\mathbf{H}^*)$ is the diagonal of the "whitened" hat matrix, $\mathbf{H}^*$.

4.16 `sprnorm()`

4.17 `vcov()`

`vcov()` returns the variance-covariance matrix of estimated parameters. Currently, `vcov()` only returns the variance-covariance matrix of the fixed effects. The variance-covariance matrix of the fixed effects is given by $(\mathbf{X}^\mathsf{T}\hat{\mathbf{\Sigma}}^{-1}\mathbf{X})^{-1}$.

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.

Goldman, Nick, Whelan, and Simon. 2000. "Statistical Tests of Gamma-Distributed Rate Heterogeneity in Models of Sequence Evolution in Phylogenetics." *Molecular Biology and Evolution* 17 (6): 975–78.

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.

Harville, David A, and Daniel R Jeske. 1992. "Mean Squared Error of Estimation or Prediction Under a General Linear Model." *Journal of the American Statistical Association* 87 (419): 724–31.

Henderson, Charles R. 1975. "Best Linear Unbiased Estimation and Prediction Under a Selection Model." *Biometrics*, 423–47.

Hoeting, Jennifer A, Richard A Davis, Andrew A Merton, and Sandra E Thompson. 2006. "Model Selection for Geostatistical Models." *Ecological Applications* 16 (1): 87–98.

Hrong-Tai Fai, Alex, and Paul L Cornelius. 1996. "Approximate F-Tests of Multiple Degree of Freedom Hypotheses in Generalized Least Squares Analyses of Unbalanced Split-Plot Experiments." *Journal of Statistical Computation and Simulation* 54 (4): 363–78.

Kackar, Raghu N, and David A Harville. 1984. "Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models." *Journal of the American Statistical Association* 79 (388): 853–62.

Kenward, Michael G, and James H Roger. 1997. "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood." *Biometrics*, 983–97.

———. 2009. "An Improved Approximation to the Precision of Fixed Effects from Restricted Maximum Likelihood." *Computational Statistics & Data Analysis* 53 (7): 2583–95.

Littell, Ramon C, George A Milliken, Walter W Stroup, Russell D Wolfinger, and Schabenberber Oliver. 2006. *SAS for Mixed Models.* SAS publishing.

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

Pinheiro, José, and Douglas Bates. 2006. *Mixed-Effects Models in S and S-Plus.* Springer science & business media.

Prasad, NG Narasimha, and Jon NK Rao. 1990. "The Estimation of the Mean Squared Error of Small-Area Estimators." *Journal of the American Statistical Association* 85 (409): 163–71.

Satterthwaite, Franklin E. 1946. "An Approximate Distribution of Estimates of Variance Components." *Biometrics Bulletin* 2 (6): 110–14.

Schluchter, Mark D, and Janet T Elashoff. 1990. "Small-Sample Adjustments to Tests with Unbalanced Repeated Measures Assuming Several Covariance Structures." *Journal of Statistical Computation and Simulation* 37 (1-2): 69–87.

Searle, Shayle R, George Casella, and Charles E McCulloch. 2009. *Variance Components.* John Wiley & Sons.

Self, Steven G, and Kung-Yee Liang. 1987. "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions." *Journal of the American Statistical Association* 82 (398): 605–10.

Stram, Daniel O, and Jae Won Lee. 1994. "Variance Components Testing in the Longitudinal Mixed Effects Model." *Biometrics*, 1171–7.

Wolf, Helmut. 1978. "The Helmert Block Method-Its Origin and Development." In *Proceedings of the Second International Symposium on Problems Related to the Redefinition of North American Geodetic Networks,(NOAA, Arlington-Va, 1978)*, 319–26.

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.