PROFILE LIKELIHOODS FOR GEOSTATSISTICAL MODELS USING GPU'S

A PREPRINT

Ruoyong Xu
Department of Statistical Sciences
University of Toronto
700 University Ave., Toronto, ON M5G 1Z5, Canada
ruoyong.xu@mail.utoronto.ca

Patrick Brown
Department of Statistical Sciences
University of Toronto
700 University Ave., Toronto, ON M5G 1Z5, Canada
patrick.brown@utoronto.ca

ABSTRACT

Profile likelihoods are rarely used in geostatistical models due to the computational burden imposed by repeated likelihood evaluations. Accounting for uncertainty in covariance parameters can be highly consequential in geostatistical models as some covariance parameters are poorly identified, the problem is severe enough that the differentiability parameter of the Matern correlation function is typically treated as fixed. The problem is compounded with anisotropic spatial models as there are two additional parameters to consider. This paper makes the following three contributions:

1- a methodology and software implementation for highly parallel computing profile likelihoods on GPU's is presented; 2- as expected, the profile-based confidence intervals have superior coverage to the more standard Wald-type intervals; and 3- a simulation study shows that despite the inconsistency inherent in maximum likelihood estimation of geostatistical models, MLE's of covariance parameters (including the shape parameter) are unbiased and credible regions have the desired coverage.

Keywords First keyword · Second keyword · More

1 Introduction

some keypoints to mention:
1, anisotropic spatial models are important but diffcult to evaluate because it consists of many covariance parameters.
2, standard approximations of spatial models (GMRF) don't allow for anisotropy (at least as implemented).
3, gpuBatchMatrix is able to compute the profile LogL for each of the parameters in linear anisotropic spatial models, for dense matrix, use no approximation methods. allow for uncertainty in covariance parameters
4, investigated likelihood-based CI's for variance parameters, wider than using information matrix based methods (gepstatsp)

1- cI for the betas for the geostapsp package
2- Ci's using profile likelihoods from gpuBatchMatrix ... second CI's will be wider
... simulation study, profile CI's have better coverage

emphasis for paper 2? - anisotropic spatial models are important
- standard approximations of spatial models (GMRF) don't allow for anisotropy (at least as implemented).
- We do dense matrix, no approximations
- also, uncertainty in covariance parameters is important in anisotropic models because there are more of them.

- we should always allow for uncertainty in covariance parameters theta = sd, nugget, range, ratio, angle, shape, BoxCox
- ... although most frequentist spatial software doesn't, assumes theta = theta hat
- anisotropic models + boxcox + matern shape means many covariance parameters (7?)
- ... important to understand uncertainty in covariance parameters, 2D profile L interesting
- ... allow for uncertainty in covariance parameters to be reflected in CI's for betas

- The easy way: - Hard way:
Likelihood-based CI's for variance parameters

We can estimate shape, range, variance
... likelihood might be fairly flat, CI's are wide
... but CI's have good coverage

2 The linear geostatistical model

We start with a review of the linear geostatistical model:

$$
\begin{aligned}
Y_i|U(s_i) &\overset{ind}{\sim} \mathrm{N}(\lambda(s_i), \tau^2), \\
\lambda(s_i) &= X(s_i)\beta + U(s_i), \\
(U(s_1), \ldots, U(s_n))^{\intercal} &\sim \mathrm{MVN}(0, \Sigma), \\
\Sigma_{ij} = \mathrm{COV}\left[U(s_i), U(s_j)\right] &= \begin{cases} \sigma^2 R(\|s_i - s_j\|/\phi; \kappa), & i \neq j. \\ \sigma^2, & i = j. \end{cases}
\end{aligned}
$$

where $Y_i : i = 1, \ldots, n$ is the observation obtained at location s_i. Given $U(s_i)$, Y_i's are mutually independently distributed with Gaussian distribution with observation variance τ^2. $X(s_i)$ is a $1 \times p$ vector of explanatory variables at location s_i, $\beta = (\beta_1, \ldots, \beta_p)^{\intercal}$ is the corresponding vector of regression parameters. $U = (U(s_1), \ldots, U(s_n))^{\intercal}$ is a zero-mean Gaussian random field with covariance matrix Σ. The covariance between locations (s_i) and (s_j) takes the above form, where $R(\cdot)$ is the correlation function, and σ^2 is the spatial variance or variability in residual variation.

Let $Y = (Y_1, \ldots, Y_n)^{\intercal}$ be a vector of observations, and $X = (X(s_1), \ldots, X(s_n))$ be an $n \times p$ design matrix, we rewrite the model in matrix format

$$Y \sim \mathrm{MVN}(X\beta, \sigma^2 R(\phi, \kappa) + \tau^2 I).$$

Write $V = R(\phi, \kappa) + \nu^2 I$, and $\nu^2 = \tau^2/\sigma^2$, then

$$Y \sim \mathrm{MVN}(X\beta, \sigma^2 V). \tag{1}$$

2.1 Matérn correlation

The powered exponential family, the Matérn family and the spherical family are the three most commonly used families of correlation functions in geostatistics. We favour the Matérn family because of its flexibility, and it is made available in our R package ***gpuBatchMatrix***. The form of Matérn correlation used in ***gpuBatchMatrix*** is

$$R(d; \phi, \kappa) = \frac{2^{\kappa-1}}{\Gamma(\kappa)} (\sqrt{8\kappa}\frac{d}{\phi})^{\kappa} K_{\kappa}(\sqrt{8\kappa}\frac{d}{\phi}), \quad \text{where} \quad \phi > 0, \kappa > 0,$$

$K_{\kappa}(\cdot)$ denotes the modified Bessel function of the second kind of order κ, κ is the shape parameter which determines the smoothness of $U(x)$. For $\kappa = 0.5$, the Matérn correlation function coincides with the exponential correlation $\exp(-d/\phi)$. When $\kappa \to \infty$, it tends to the Gaussian correlation $\exp\{-2(\|d\|/\phi)^2\}$. ϕ is called the range or scale parameter, it controls the rate of decay of the correlation as d increases.

2.2 Box-Cox transformation

The fit of the linear geostatistical model can often be improved by applying a transformation to the response variable Y. Skewed random fields that mildly deviate from the Gaussian distribution can be modeled by means of the Box-Cox

transformation (BCT) [1]. For positive-valued response variable Y, the Box-Cox transformation has the form

$$Y' = \begin{cases} (Y^\lambda - 1)/\lambda : & \lambda \in \mathbb{R} \quad \text{and} \quad \lambda \neq 0, \\ \log Y : & \lambda = 0, \end{cases}$$

where Y' denotes the transformed response. The aim of the BCT is to make data closely resemble the Gaussian distribution, so that the usual assumptions for linear models hold. The BCT cannot guarantee that any probability distribution can be normalized. However, even in cases when an exact mapping to the normal distribution is not possible, BCT may still provide a useful approximation [2].

2.3 Geometric anisotroy

Geometrical anisotropy is defined by two additional parameters in the correlation function. Let (x_1, x_2) represent the original coordinates, which is rotated counterclockwise by φ, then the new coordinates (x_1', x_2') is given by

$$\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} = \begin{pmatrix} 1/\phi_1 & 0 \\ 0 & 1/\phi_2 \end{pmatrix} \cdot \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

where $\varphi \in [-\pi/2, \pi/2]$ is called the anisotropy angle, $\psi = \phi_1/\phi_2 \in [0, 1]$ is the anisotropy ratio.

3 Inference based on the likelihood function

Let $y'^\intercal = (y_1', \ldots, y_n')$ be the Box-Cox transformed data of the observations $y^\intercal = (y_1, \ldots, y_n)$. The most general log-likelihood function for the linear geostatistical model parameters given y' would be

$$-2\ell(\beta, \sigma^2, \lambda, \omega; y') = (y' - X\beta)^\intercal (\sigma^2 V)^{-1}(y' - X\beta) + \log|\sigma^2 V| - 2 * (\lambda - 1)\sum_{i=1}^{n} \log y_i + n\log(2\pi). \tag{2}$$

the third term on the right-hand side of (2) arises from the Jacobian of the transformation. This log-likelihood function is complex to evaluate as we have several types of parameters to estimate, including the regression parameters β, the transformation parameter λ, and the covariance parameters $(\sigma^2, \phi, \kappa, \nu^2, \varphi, \psi)$. For simplicity we use $\omega = (\phi, \kappa, \nu^2, \varphi, \psi)$ to denote the correlation parameters in the following.

3.1 Profile log-likelihood functions

The profile log-likelihood for parameter φ is defined as $\ell_p(\varphi; y) = \sup_{\psi} \ell(\varphi, \psi; y)$. Given ω and λ, The log-likelihood function (2) is maximized at

$$\hat{\beta}_{\omega,\lambda} = (X^\intercal V^{-1} X)^{-1} X^\intercal V^{-1} y', \quad \text{and} \tag{3}$$

$$\hat{\sigma}^2_{\omega,\lambda} = \frac{1}{n}(y' - X\hat{\beta}_{\omega,\lambda})^\intercal V^{-1}(y' - X\hat{\beta}_{\omega,\lambda}). \tag{4}$$

Substitute (3) and (4) back into (2), we have the profile log-likelihood for ω and λ

$$-2\ell_p(\omega, \lambda; y') = n\log\frac{(y' - X\hat{\beta}_{\omega,\lambda})^\intercal V^{-1}(y' - X\hat{\beta}_{\omega,\lambda})}{n} + \log|V| - 2(\lambda - 1)\sum_{i=1}^{n}\log y_i + n\log(2\pi) + n, \tag{5}$$

The main problem of maximum likelihood estimation is that variance is underestimated. Think of the simple case when the covariance matrix of Y is $\sigma^2 I$, then

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^\intercal (y - X\hat{\beta}),$$
$$\mathbb{E}[\hat{\sigma}^2] = \frac{n-p}{n}\sigma^2 < \sigma^2,$$

where $p = \text{rank}(X)$, is the number of elements in β. The estimation bias comes from not taking into account the degree of freedom used for estimating β. Restricted maximum likelihood (REML) developed by [3] is an approach that produces unbiased estimators or less biased estimators than MLE in general. REML eliminates the influence of X on $\hat{\sigma}^2$ and $\hat{\omega}$. In this method, we find a matrix A of full rank such that $AX = 0$. The $n \times n$ projection matrix

$S = I - X(X^\intercal X)^{-1}X^\intercal$ satisfies $SX = 0$, however, S is degenerate as rank$(S) = n - p$, so we take just $n - p$ linearly independent rows from S to make the matrix $A_{(n-p)\times n}$. Transform the data linearly to $Y^* = AY$, then $\mathbb{E}(Y^*) = AX\beta = 0$. Recall in (1) we obtained $Y \sim \text{MVN}(X\beta, \sigma^2 V)$, thus

$$Y^* \sim \text{MVN}(0, \sigma^2 AVA^\intercal).$$

The principle of the REML method is to estimate the variance parameter σ^2 by maximizing the restricted log-likelihood function below, where y^* represents a realization of Y^*,

$$-2\ell^*(\sigma^2, \omega; y^*) = y^{*\intercal}(\sigma^2 AVA^\intercal)^{-1}y^* + \log|\sigma^2 AVA^\intercal| + (n-p)\log(2\pi), \tag{6}$$

Differentiating (6) with regard to σ^2 and setting it equal to zero gives the unbiased estimate for σ^2

$$\hat{\sigma}^2_{reml}(\omega) = \frac{y^{*\intercal}(AVA^\intercal)^{-1}y^*}{n-p}.$$

The REML criterion is based on the likelihood of Y^*, in which X does not appear. REML bypasses estimating β and can therefore, produce unbiased estimates for σ^2.

We can write (6) in terms of the original observed data y using the two identities derived by [4] in Section 5.1:

$$|A(\sigma^2 V)A^\intercal| = |\sigma^2 V||X^\intercal(\sigma^2 V)^{-1}X|$$

and

$$y^*(\sigma^2 AVA^\intercal)^{-1}y^* = (y - X\hat{\beta})^\intercal(\sigma^2 V)^{-1}(y - X\hat{\beta}),$$

where $\hat{\beta} = (X^\intercal V^{-1}X)^{-1}X^\intercal V^{-1}y$. Then the restricted log-likelihood of y is

$$-2\ell^*(\sigma^2, \omega; y) = (y - X\hat{\beta})^\intercal(\sigma^2 V)^{-1}(y - X\hat{\beta}) + (n-p)\log\sigma^2 + \log|V| + \log|X^\intercal V^{-1}X| + n\log(2\pi).$$

Applying the BCT on y we have the restricted log-likelihood in terms of the transformed response data y'

$$-2\ell^*(\sigma^2, \omega, \lambda; y') = (y' - X\hat{\beta})^\intercal(\sigma^2 V)^{-1}(y' - X\hat{\beta}) + (n-p)\log\sigma^2 + \log|V| + \log|X^\intercal V^{-1}X| + n\log(2\pi) - 2(\lambda - 1)\sum_{i=1}^{n}\log y_i, \tag{7}$$

which does not depend on the choice of matrix A.

Differentiate (7) with respect to σ^2 and set it equal to zero, we have the expression for $\hat{\sigma}^2_{reml}$ in terms of the Box-Cox transformed data given ω and λ,

$$\hat{\sigma}^2_{reml}(\omega, \lambda) = \frac{(y' - X\hat{\beta})^\intercal V^{-1}(y' - X\hat{\beta})}{n-p}. \tag{8}$$

Substitute (8) into (7), leading to the profile restricted log-likelihood for (ω, λ)

$$-2\ell^*_p(\omega, \lambda; y') = (n-p)\log\hat{\sigma}^2 + \log|V| + \log|X^\intercal V^{-1}X| - 2(\lambda - 1)\sum_{i=1}^{n}\log y_i + n\log(2\pi) + n - p. \tag{9}$$

Compared to the "ml" profile log-likelihood given in (5), the effective dimension is reduced from n of $\ell(\cdot)$ to $(n-p)$ of $\ell^*_p(\cdot)$, this distinction is important if p is large.

3.2 Parameter estimation

We write $I(\theta)$ for the expected Fisher information, here θ represents the true parameter value. $I(\hat{\theta})$ is the observed Fisher information evaluated at the MLE of θ.

Wald confidence interval $\{\hat{\theta} \pm z_{1-\alpha/2} * I(\hat{\theta})^{-1/2}\}$ is a $100(1-\alpha)\%$ confidence interval for θ, where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ percentile of the standard normal distribution.

Likelihood-based confidence interval $\{\theta : \ell(\theta) \geq \ell(\hat{\theta}) - c_p(\beta)/2\}$ is asymptotically, a β-level confidence set for θ, where $c_p(\beta)$ is the β-quantile of χ^2_p, i.e., $\mathbb{P}(\chi^2_p \leq c_p(\beta)) = \beta$.

Wald intervals are always symmetric, and computationally much easier than likelihood-based confidence intervals, as quantities need to be calculated are all available from the algorithm used to find the MLEs of the model parameters. However, Wald intervals may include values which are not valid for the parameter in question. And if the likelihood function is not regular (cannot be well approximated by a quadratic function), then the curvature of the log-likelihood at the MLE ($I(\hat{\theta})$) or the standard error is not meaningful, which is often the case for spatial models where the parameter estimate is near a boundary, and the likelihood function is quite flat. In this case, Wald interval is deficient, likelihood ratio based confidence interval is a better and safer choice [5]. For that reason, plotting the profile likelihood function in unfamiliar problems is generally advised. "Further, a situation in which the Wald approach completely fails while the likelihood ratio approach is still (often) reasonable is when testing whether a parameter lies on the boundary of its parameter space"

Geostatsp's methods

The *lgm()* function in ***geostatsp*** package estimates the correlation parameters $\omega = (\phi, \kappa, \nu^2, \varphi, \psi)$ and the transformation parameter λ by maximization of the likelihood. Specifically, *lgm()* uses the *optim* function to optimize the equation (5) (if *reml()*=FALSE) numerically over many different sets of (ω, λ) to find $(\hat{\omega}, \hat{\lambda}) = \text{argmax}_{\omega,\lambda}\, \ell_p(\omega, \lambda; y')$. ***geostatsp*** then treats $\hat{\omega}$ and $\hat{\lambda}$ as fixed known parameters and back substitute them into (3) and (4) to obtain the MLEs $\hat{\beta}_{\omega,\lambda}$ and $\hat{\sigma}^2_{\omega,\lambda}$. ***geostatsp*** calculates Wald confidence intervals for model parameters in *lgm()*. The *numderiv()* function inside *lgm()* returns the Hessian matrix evaluated at $\hat{\omega}$ and $\hat{\lambda}$, the diagonals of the inverse of the negative Hessian is used for approximating the variance of the parameters.This way does not take into account of uncertainty in correlation parameters. ***geostatsp*** does the above computation process in C, and (ADD LATER!) which has achieved a great perfomance in terms of speed.

The other function *profLlgm()* in ***geostatsp*** calculates Likelihood-based confidence intervals. *profLlgm()* calculates the profile log-likelihood for each of the correlation parameter ω and λ, specifically, for estimating for example the range parameter ϕ, for a fixed value of ϕ, it uses *optim()* to maximize (5) over a number of different combinations of $(\kappa, \nu^2, \varphi, \psi)$, and repeat this process for a number of ϕ values, and use *approxfun()* to get curve of the profile log-likelihood and MLE of ϕ, then computes the confidence interval based on the profile log-likelihood. *profLlgm()* runs *optim()* a great many of times and can be time-consuming.

gpuBatchMatrix's method

The hard way is by computing the profile log-likelihood $\ell_p(\beta) = \ell(\beta, \hat{\omega}_\beta, \hat{\sigma}_\beta, \hat{\lambda}_\beta)$. (QUESTION: cannot geostatsp uses optim for beta? is it too slow so geostatsp does not do that?)

SHOW A PROFILE LOG-LIKELIHOOD PLOT FOR BETA'S. (MEUSE DATA OR SWISSRAIN DATA?)
SHOW ALL LIKELIHOODS PLOT VS BETA'S

"The results R1 and R2 are invoked routinely in modern statistical practice, but they rely on assumptions that are not always satisfied. In a geostatistical context, the two most common situations in which these assumptions do not hold are boundary problems and dimensional ambiguities. "

3.3 CI's comparison for model parameters

SHOW A TABLE OF CI compare between ***geostatsp*** and ***gpuBatchMatrix***

```
R> library(gpuRandom)
R> library(gpuR)
R> library(data.table)
R> # library(geostatsp)
```

4 A simulation study

```
R> count/Nsim
## [1] 0.84
R> mean_estimates
## (Intercept)        cov1        cov2   sdSpatial       range
```

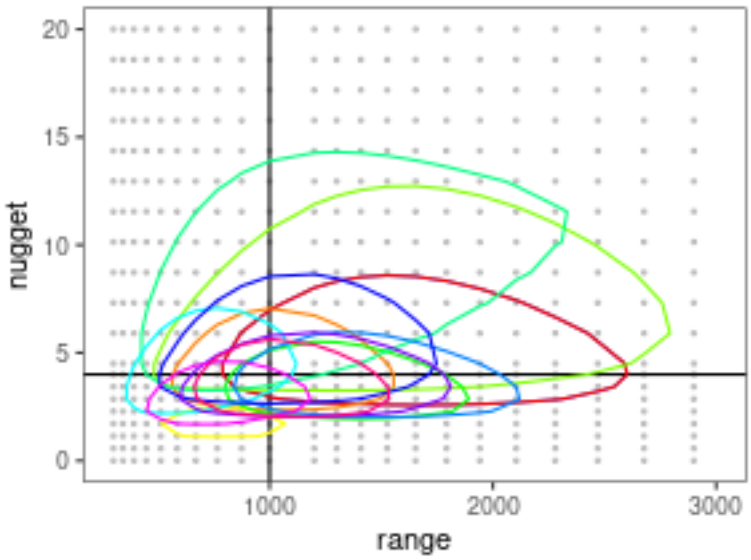

Figure 1: title

```
##       3.119       1.000       0.499       0.510    1006.381
##      nugget
##       4.084
```

5 Discussion

References

[1] George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.

[2] Dionissios T Hristopulos. *Random fields for spatial data modeling: a primer for scientists and engineers*. Springer Nature, 2020.

[3] H Desmond Patterson and Robin Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

[4] SR Searle, RL Quaas, et al. A notebook on variance components: A detailed description of recent methods of estimating variance components, with applications in animal breeding. 1978.

[5] Yudi Pawitan. *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, 2001.