Approximating likelihoods for large spatial data sets

By Michael Stein, Zhiyi Chi, and Leah Welty

Jessi Cisewski

April 28, 2009

1 Introduction

Given a set of spatial data, often the desire is to estimate its covariance structure. For practical purposes, it is often necessary to propose some parametric model for the variagram. Once the parametric model is selected, one must estimate the parameters of the model. Given a parametric structure, it naturally follows that maximum likelihood or Bayesian methods can be used in estimation. There are also nonparametric methods available, but the focus of Stein et al. [1] is on likelihood-based methods for parameter estimation of the covariance structure of a Gaussian random field. The issue with likelihood-based methods is that when the observations are irregularly sited, the computational effort can be large. Even for Gaussian random fields, the computational effort is $O(n^3)$ for each calculation (Stein et al.).

Vecchia (1988) [2] proposed a method to approximate the likelihood of spatial data sets using the basic fact that any joint density can be written in terms of the product of conditional densities. Let $\mathbf{Z} = (Z_1, \cdots, Z_n)'$ have joint density $p(\mathbf{z}; \phi)$, where p is a generic density, and ϕ is an unknown vector-valued parameter.
Partition $\mathbf{Z}$ into subvectors $\mathbf{Z_1}, \cdots, \mathbf{Z_b}$ with $\mathbf{Z}'_{(\mathbf{j})} = (\mathbf{Z}'_\mathbf{1} \cdots \mathbf{Z}'_\mathbf{j})$. Then,

$$p(\mathbf{z}; \phi) = p(\mathbf{z_1}; \phi) \prod_{j=2}^{b} p(\mathbf{z_j} | \mathbf{z_{(j-1)}}; \phi) \tag{1}$$

Vecchia (1988) revised the conditioning vectors for $j = 1, \cdots, b-1$, to a subvector, $\mathbf{S_{(j)}}$, of $\mathbf{Z_{(j)}}$, then

$$p(\mathbf{z}; \phi) \approx p(\mathbf{z_1}; \phi) \prod_{j=2}^{b} p(\mathbf{z_j} | \mathbf{s_{(j-1)}}; \phi).$$

Stein et al. develops this approximation for restricted maximum likelihood (REML), which is used in estimation of the parameters of the covariance structure. REML is preferred

over ML for this type of estimation because it does not rely on estimation of the mean function. Standard ML methods use the estimated mean parameters in the estimation of the covariance parameters, but acts as if the mean parameters are known. Since the standard ML method does not account for the uncertainty in the estimation of the mean parameters, this leads to underestimation of the variance.

2 Methods

The methods developed in this paper are for Gaussian random fields, and so the general framework is based on the supposition that $\mathbf{Z} \sim N\{F\beta, K(\theta)\}$. Furthermore, F is considered to be a known $n \times p$ matrix of rank p, and β is a p-dimensional vector of unknown coefficients. The focus of the paper is concerned with estimation the parameters, θ, of the covariance matrix K. The assumptions are that the covariance matrix K is positive definite for all $\theta \in \Theta$, the vector $\mathbf{Z}_i$ has length n_i, and the rank of F_1 is p. In addition, we assume that the best linear unbiased predictor (BLUP) of $\mathbf{Z}_j$ in terms of $\mathbf{Z}_{(j-1)}$ exists for all $\theta \in \Theta$.

In the methodology section, they first develop the restricted likelihood form in terms of the errors of the BLUPs of $\mathbf{Z}_j$ based on $\mathbf{Z}_{(j-1)}$, where $\mathbf{Z}_{(j-1)} = \mathbf{Z}_1 + \cdots + \mathbf{Z}_{(j-1)}$. $\mathbf{W}_j(\theta) = B_j(\theta)\mathbf{Z}$, where $B_j(\theta)$ is the $n_j \times n$ matrix such that $\mathbf{W}_j(\theta)$ is the vector of the errors of the BLUP of $\mathbf{Z}_j$ in terms of $\mathbf{Z}_{(j-1)}$. The result is the following proposition.

Proposition 1. The restricted log-likelihood of θ in terms of $\mathbf{Z}$ is given by

$$rl(\theta; \mathbf{Z}) = -\frac{n-p}{2} log(2\pi) - \frac{1}{2} \sum_{j=1}^{b} [log\{det(V_j)\} + \mathbf{W}_j' V_j^{-1} \mathbf{W}_j].$$

The approximation of the restricted maximum likelihood displayed in Proposition 1 has the same form, but $\mathbf{W}_j$ for $j > 1$ is the error of the BLUP of $\mathbf{Z}_j$ based on the sub-vector $\mathbf{S}_{(j-1)}$ of $\mathbf{Z}_{(j-1)}$.

Now that the form of the approximation to the restricted log-likelihood has been developed, there are several issues to address regarding the use of this approximation. The length of the prediction vectors ($\mathbf{Z}_j$) and conditioning vectors ($\mathbf{S}_{(j-1)}$) must be decided, along with which elements of $\mathbf{Z}_{(j-1)}$ should be included in $\mathbf{S}_{(j-1)}$. Vecchia (1998) noted, and Stein et al. agrees that any statistical method used to make these decisions would be based on unknown parameters. Fortunately, Stein et al. came up with a way to check the efficiency of the approximation by using the well-developed theory of estimating equations. An explanation of this validation method is explained in the next section covering the numerical results.

3 Numerical Results

As noted in the paper, there are manifold possible networks and designs that could be studied in order to observe the effects of the approximated restricted likelihood on covariance structure parameter estimation under prediction and conditioning vectors of varying lengths. They decided to consider an observation network composed of 1000 randomly selected sites (out of 10,000 points in a plane). Each selected point is then perturbed by adding a random point in $[0.25, 0.25]^2$. In the first cases considered, they include only one observation in the prediction vector $\mathbf{Z}_j$ for $j > 1$. In the paper, they also consider designs with longer prediction vectors.

In these designs, they allow the conditioning vectors to be lengths $m = 8, 16$, and 32, and consider three cases regarding the distance from the elements of the conditioning vector to the prediction vector: all m nearest neighbors, three-quarters of the m as nearest neighbors, and half of the m as nearest neighbors . The covariance structures for the random fields are the exponential model ($cov\{Z(\mathbf{x}), Z(\mathbf{y})\} = \theta_2 exp(-\theta_1|\mathbf{x} - \mathbf{y}|/\theta_2)$) and the power law variogram model ($\frac{1}{2}var\{Z(\mathbf{x}) - Z(\mathbf{y})\} = \theta_2|\mathbf{x} - \mathbf{y}|^{\theta_1}$).

I have included the table of the relative efficiencies for the exponential model (Table 1). I include the table to clarify the design and procedure developed in the paper. The summary of results for the power law variogram model has the same form (except that the efficiencies for a non-constant mean function is also displayed), but I direct the reader to the paper to get the numerical results. The results displayed in Table 1 reflect the efficiency of the parameter estimation under the approximation to the restricted likelihood. As alluded to in the methodology section, using estimating equations leads to what is referred to as the robust information measure. This robust information measure provides a method for checking the efficiency of the approximation to the restricted likelihood in estimation compared to the exact restricted likelihood. It is shown that if there is no restriction on the conditioning vector $\mathbf{S}_{(j-1)}$, then we can take $\mathbf{S}_{(j-1)} = \mathbf{Z}_{(j-1)}$, in which case the robust information measure is the Fisher information matrix. Hence the robust information measure based on the approximation can be thought of as a reflection of the extent to which information is lost by using the approximation. The numbers displayed in Table 1 are the ratios displayed as percentages of the robust information measures to the Fisher information matrix. Note that the Fisher information matrix can be calculated because they consider only 1000 observations. Numbers close to 100 percent reflect more efficient estimation.

As expected, designs with longer conditioning vectors performed better than designs with shorter conditioning vectors. What is important to note in the tables of results in the paper are which portion of nearest neighbors to distant observations do the best, and under which lengths of conditioning vectors. For example, for Table 1 we see that component 1 performs similarly under the various designs given the conditioning vector length. However, component 2's performance highly depends on the particular design. The general interpretation for component 2 is that when component 1 is smaller ($\theta_1 = 0.02$ or 0.1), it

does not perform well when the conditioning vector is composed only of nearest neighbors; however, for larger values of component 1 ($\theta_1 = 0.5$ or 2), it performs competitively or even better when only nearest neighbors are considered. There are many more interesting results of this nature included in the paper.

Table 1. Relative efficiencies of estimators by using approximate restricted likelihoods compared with the exact REML estimators as measured by the diagonal elements of the inverse information matrices based on 1000 observations as described in Section 4.1†

θ_1	m'/m	*Relative efficiencies (%) for the following components and values of m:*					
		Component 1			*Component 2*		
		$m=8$	$m=16$	$m=32$	$m=8$	$m=16$	$m=32$
0.02	1	87.3	94.8	97.5	24.8	33.6	44.7
	0.75	91.1	97.4	99.2	71.0	83.1	90.5
	0.5	85.7	95.2	98.6	74.7	82.9	90.9
0.1	1	83.1	91.6	95.3	35.3	48.9	61.9
	0.75	89.0	96.6	99.0	58.6	78.4	89.8
	0.5	88.0	95.8	98.9	72.9	85.8	93.7
0.5	1	79.7	89.1	94.2	63.1	78.9	88.4
	0.75	77.6	89.3	94.4	58.2	78.2	88.8
	0.5	81.9	92.2	96.2	69.1	81.6	91.0
2	1	84.2	91.6	95.6	88.6	94.1	97.0
	0.75	81.8	89.8	94.6	83.9	92.1	96.0
	0.5	79.8	88.8	94.1	80.5	90.4	95.2

†The designs that are used for approximate likelihoods are of the form $D(m, m')$ as defined in Section 4.2. The model for the covariance function is $\theta_2 \exp(-\theta_1 d/\theta_2)$, where d is the interpoint distance; $\theta_2 = 1$ in all cases.

4 Conclusion

While this paper develops a number of interesting ideas, and addresses many more concepts than I include in my presentation or this summary, the authors note that there is a lot of room for further development of these ideas. Vecchia (1988) came up with a clever way to approximate the likelihood of large spatial data sets, thus allowing for feasible likelihood-based methods for large spatial data sets. Stein et al. made a number of extensions of these methods, including applying the approximation to REML, and developed a method to verify the efficiency of these approximations under various designs using estimating equations and the robust information measure. Stein et al. also discusses at several points in the paper that designs that are good for estimation are not necessarily good for

prediction, and vice versa. Hence, the goal of estimation or prediction must be considered when determining the design.

References

[1] Stein, M. L., Chi, Z. and Welty, L. J. (2004) Approximating likelihoods for large spatial data sets. *J. R. Statistic. Soc.* **B, 66**, 275-296.

[2] Vecchia, A.V. (1988) Estimation and model identification for continuous spatial processes. *J. R. Statistic. Soc.* **B, 50**, 297 - 312.