GAUSSIAN PROCESSES FOR LARGE SPATIAL DATA

GAL AV-GAY - 48300057

1. Overview

Modelling spatial data with gaussian processes is a common approach for geo-statistical analyses. Typically, it is assumed that spatially referenced observations are multivariate-normally distributed with a covariance matrix that depends on unknown parameters θ.

$$y \sim N(X\beta, R(\theta)) \tag{1}$$

Where the covariance matrix $R(\theta)$ defines the correlation between any two given observations. This covariance matrix completely specifies the gaussian process, as the regression parameters $\hat{\beta}$ are a function of $R(\theta)$ as such,

$$\hat{\beta} = (F^T R^{-1} F)^{-1} F^T R^{-1} y$$

where F is a vector of the regression functions, and is in the constant case an $n \times 1$ vector of 1s.

A number of different correlation functions are commonly used for defining the values in this covariance matrix $R(\theta)$, some of which will be discussed below. Inference on these θ parameters changes depending on the correlation function used. It is common for a correlation function to assume that points that are closer together have a higher correlation, for example in the gaussian correlation function the euclidean distance is used:

$$R(x, x^{'}) = R(\theta) = \prod_{j=1}^{k} exp\left[-\theta_j |x_j - x_j^{'}|^2\right] \tag{2}$$

Inference on these theta parameters in this case requires maximization of the following log-likelihood,

$$-\frac{n}{2} ln(\hat{\sigma}^2(\mathbf{R})) - \frac{1}{2} ln|\mathbf{R}| + const \tag{3}$$

Evaluation of this log-likelihood, as well as derivatives of this likelihood, involves taking the inverse and determinant of the covariance matrix $\mathbf{R}$. This is implemented using the cholesky decomposition, which requires n^3 floating point operations, where n is the number of observations. This cost is exacerbated by the use of iterative procedures for estimation of the θ parameters.

Prediction is also limited by this computational complexity. Given a

Date: Dec 12th 2014.

gaussian process $z(s)$, we wish to predict $z(x^*)$ given the n observations and a new $x^* \in \mathcal{D}$. To do this we take

$$\hat{z}(x^*) = r^{*'} R^{-1} Z \tag{4}$$

where $Z = (Z(x_1), ..., Z(x_n))'$, and $r_i^* = R(x_i, x^*)$. In this case the best linear unbiased predictor (BLUP) can be written as $\tilde{Z(x^*)} = r^{*'} u$, with $Ru = Z$. Solving this system of linear equations to find the inverse of the covariance matrix is done most efficiently using the cholesky decomposition.

The methods discussed in this manuscript are primarily concerned with lowering the computational cost of maximizing this log likelihood for large n. These methods achieve improved computational efficiency by dimensionality reduction via "low-rank" models or by assuming sparsity.

2. Low-Rank models

Low-Rank models are representations of the spatial process on a lower dimensional subspace, where a subset of the n original locations $r < n$ is considered. These processes result in a reduced computational complexity of $O(nr^2 + r^3) \approx O(nr^2)$.

One such representation involves dimensionality reduction of a spatial process that is specified by the convolution of i.i.d random variables (Higdon 2001). There are a number of benefits to this formulation aside from dimensionality reduction. Features such as non-stationarity, edge effects, non-Gaussian fields, and alternative space-time models can easily be accommodated using this formulation. A gaussian process $z(s)$ over spatial region S can be constructed by convolving a continuous latent model $x(s), s \in S$ with smoothing kernel $k(s)$, such that

$$z(s) = \int_s k(u - s)(u) du$$

For all $s \in S$. The resulting covariance function for this process depends only on the displacement vector $d = s - s'$.

$$c(d) = Cov(z(s), z(s')) = \int_S k(u - d)k(u) du$$

Based on the convolution theorem for fourier transforms, Higdon shows that there exists a one to one relationship between the smoothing kernel and the covariance matrix, given that $\int_{\mathbb{R}^p} k(s) ds < \infty$ and $\int_{\mathbb{R}^p} k^2(s) ds < \infty$, where $k(s)$ is the smoothing kernel, or given that the covariance matrix is integrable and positive definite.

Dimensionality reduction is achieved by restricting the latent process $x(s)$ to locations $s_1, ..., s_m$ where $m < n$. In this case, a small number of parameters $x(s_1), ..., x(s_m)$ control the entire spatial process. The covariance function is determined by the latent process and the smoothing kernel,

$$z(s) = \sum_{j=1}^{m} x_j k(s - z_j)$$

The disadvantages to this formulation are shared with other methods that involve dimensionality reduction: When n is large, simulation studies suggest that m must be fairly large to adequately approximate the full model[2]. In addition, these models result in poor likelihood approximations when neighbouring observations are strongly correlated and the spatial signal is far greater than the noise.

3. Sparse-Covariance Methods

Sparse-covariance methods assume that the correlation between non-neighbouring observations is effectively zero and therefore distant observations should be considered independent. One example of a sparse-covariance method is the method of covariance-tapering, as described in Furrer et. al. (2006).

Consider a gaussian process $z(s)$ with covariance function $R(s, s^{'})$, where $s \in \mathcal{D} \subset \mathbb{R}^d$, observed at n locations $s_1, ..., s_n$. Covariance tapering deliberately introduces zeroes into the covariance matrix to make it sparse, in such a way that the positive definiteness of the matrix is maintained. Naive predictions can then be obtained by replacing R and r in (3) and (4) with the new, tapered, positive definite covariance matrix, R_{tap}.

The effect of 'misspecifying' the covariance function is covered in a number of articles by Stein (1988, 1990, 1993), which investigate the asymptotic optimality and efficiency of an incorrect covariance function. Application of this theory is reinterpreted in Furrer et. al. in order to evaluate the increase in squared prediction error that results from a deliberately modified covariance as in the covariance tapering method. Using spectral densities, Furrer et. al. prove the asymptotic equivalence of the original predictor in (4) and the tapered predictor for a Matern covariance function.

The advantage of this model over other sparse models is that it offers fully process based inference.

4. Nearest-Neighbours Gaussian Processes

Limiting the covariance function to a local neighbourhood is not a new idea. There are a number of ways to achieve this. One way described by Gribov and Krivoruchko (2004) implements covariance tapering using a moving neighbourhood to generate continuous prediction and prediction standard error surfaces. In this case the tapering is dependent on the prediction and the data location.

Datta et. al. (2014) claims that none of the previously mentioned spatial processes unify estimation, prediction, and model assessment. This claim is used as motivation for the construction of a well-defined spatial process named Nearest-Neighbor Gaussian Process (NNGP). Based on work in Vecchia (1988) and Stein et. al. (2004) that shows how likelihood approximations using lower-dimensional conditional distributions result in proper densities under general conditions, Datta. et. al. show that these densities

are distributed according to finite-dimensional realizations of an NNGP. This well-defined process adopts a Bayesian modelling framework using posterior predictive distributions in order to demonstrate its inferential capabilities. Furthermore, a computationally efficient Gibbs sampling algorithm is developed to conduct inference.

Below we summarize the formulation of the NNGP that provides the basis of the framework for Bayesian estimation and prediction with this flexible setup.

Given $z(s) \sim GP(0, R(\cdot,\cdot|\theta))$, a zero-centred q-variate Gaussian process, where $z(s) \in \mathbb{R}^{q\times 1}$ for all $s \in \mathcal{D} \subset \mathbb{R}^d$, the cross covariance function $R(\cdot,\cdot|\theta)$ maps a pair of locations s and t in $\mathcal{D} \times \mathcal{D}$ into a $q \times q$ real valued matrix $C(s,t)$ with entries $cov\{w_i(s), w_j(t)\}$. Let $S = \{s_1, ..., s_k\}$ be a fixed finite collection of locations in $\mathcal{D}$, called the reference set, and z_A denotes the realizations of the gaussian process $z(s)$ over a finite collection of locations A. This gives $z_s \sim N(0, C_s(\theta))$, where $z_s = (z(s_1)', z(s_2)', ..., z(s_k)')'$ and $C_s(\theta)$ is a positive definite $qk \times qk$ black matrix with $C(s_i, s_j)$ as its blocks. The joint density of z is written as $p(z) = \prod_{i=1}^{k} p(z(s_i)|z_{<i})$, where $z_{<i} = (z(s_1)', z(s_2)', ..., z(s_{i-1})')'$.

Dimensionality reduction is achieved when instead of considering $z_{<i}$, a smaller set $N(s_i)$ is defined for every $s_i \in S$, consisting of neighbours of s_i. The following composite likelihood is obtained:

$$\tilde{p}(z_s) = \prod_{i=1}^{k} p(z(s_i)|z_{N(s_i)}) \tag{5}$$

Datta et. al. prove that if we view the pair $\{S, N_S\}$ as a directed graph $\mathcal{G}$ with vertices at $s_1, ..., s_k$, and directed edges to every vertex s_i from all locations in $N(s_i)$, then $\tilde{p}$ in (5) is a valid joint distribution. Furthermore, given that $N(s_i)$ identifies the m nearest neighbours from the past, ensuring an acyclic $\mathcal{G}$, it is proven that $\tilde{p}$ is a Gaussian density with a sparse precision matrix. Datta et. al. proceeds to specify a full posterior distribution given priors on β, θ, and τ_j^2, where τ_j^2 are the diagonal elements of a dispersion matrix D that specifies the covariance of the random errors for each observation, $\epsilon(t) \sim^{i.i.d} N(0, D)$. This posterior is used for an implementation of Gibbs algorithm, which facilitates inference at a reduced computational complexity.

5. Conclusion

We have highlighted a number of methods for curbing the computational cost of doing inference, prediction, and model assessment, for a Gaussian process on a large spatial data-set, as well as the advantages and disadvantages of specific methods. The research in this area is quite rich, as there are multiple different approaches in both the frequentist and bayesian interpretations. The Datta et. al. offer a holistic bayesian approach that

facilitates fully model-based prediction and inference, as well as model assessment. This approach is optimal in some context, however, other methods offer practical benefits and features that may be preferred depending on the application or context of the problem being solved.

6. BIBLIOGRAPHY

(1) Higdon, D. (2001), Space and Space Time Modeling using Process Convolutions, *Technical Report, Institute of Statistics and Decision Sciences, Duke University, Durham.*

(2) Stein, M. L. (2013), Limitations on Low Rank Approximations for Covariance Matrices of Spatial Data, *Spatial Statistics.*

(3) Furrer, R., Genton, M. G., and Nychka, D. (2006), Covariance Tapering for Interpolation of Large Spatial Datasets, *Journal of Computational and Graphical Statistics*, 15, 503-523.

(4) Stein, M. L. (1988), Asymptotically Efficient Prediction of a Random Field with a Misspecified Covariance Function, *The Annals of Statistics*, 16, 5563.

(5) Stein, M. L. (1990b), Uniform Asymptotic Optimality of Linear Predictions of a Random Field Using an Incorrect Second-Order Structure, *The Annals of Statistics*, 18, 850-872.

(6) Stein, M. L. (1993), A Simple Condition for Asymptotic Optimality of Linear Predictions of Random Fields, *Statistics & Probability Letters*, 17, 399-404.

(7) Gribov, A., and Krivoruchko, K. (2004), Geostatistical Mapping with Continuous Moving Neighborhood, *Mathematical Geology*, 36, 267-281.

(8) Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2014). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. *Tech. rep., Univeristy of Minnesota.*

(9) Stein, M. L., Chi, Z., and Welty, L. J. (2004), Approximating Likelihoods for Large Spatial Data Sets, *Journal of the Royal Statistical Society. Series B (Methodological)*, 66, 275- 296.

(10) Vecchia, A. V. (1988), Estimation and Model Identification for Continuous Spatial Processes, *Journal of the Royal Statistical Society. Series B (Methodological)*, 50, 297-312.