Computational Statistics and Data Analysis 123 (2018) 116–130

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A sparse linear algebra algorithm for fast computation of prediction variances with Gaussian Markov random fields ☆

Andrew Zammit-Mangion [a,*], Jonathan Rougier [b]

[a] National Institute for Applied Statistics Research Australia (NIASRA), School of Mathematics and Applied Statistics (SMAS), University of Wollongong, Northfields Avenue, Wollongong, NSW 2522, Australia
[b] School of Mathematics, University of Bristol, Tyndall Avenue, Bristol, BS8 1TH, UK

ARTICLE INFO

Article history:
Received 18 July 2017
Received in revised form 16 December 2017
Accepted 5 February 2018
Available online 15 February 2018

Keywords:
Conditional dependence
GMRF
Lattice spatial model
Sparse inverse subset
Takahashi equations

ABSTRACT

Gaussian Markov random fields are used in a large number of disciplines in machine vision and spatial statistics. The models take advantage of sparsity in matrices introduced through the Markov assumptions, and all operations in inference and prediction use sparse linear algebra operations that scale well with dimensionality. Yet, for very high-dimensional models, exact computation of predictive variances of linear combinations of variables is generally computationally prohibitive, and approximate methods (generally interpolation or conditional simulation) are typically used instead. A set of conditions is established under which the variances of linear combinations of random variables can be computed exactly using the Takahashi recursions. The ensuing computational simplification has wide applicability and may be used to enhance several software packages where model fitting is seated in a maximum-likelihood framework. The resulting algorithm is ideal for use in a variety of spatial statistical applications, including *LatticeKrig* modelling, statistical downscaling, and fixed rank kriging. It can compute hundreds of thousands exact predictive variances of linear combinations on a standard desktop with ease, even when large spatial GMRF models are used.

1. Introduction

Gaussian Markov random fields (GMRFs) play a pivotal role in various applications such as image analysis (Mardia, 1988), disease mapping (Lawson and MacNab, 2011), and atmospheric pollution modelling (Cameletti et al., 2013). They are frequently seen as reasonable approximations to continuously-indexed Gaussian processes (Rue and Tjelmeland, 2002), and are often preferred due to their favourable computational properties. Recent work on their ability to approximate Gaussian processes typically used in geostatistical models (e.g., Lindgren et al., 2011; Nychka et al., 2015) has led to their widespread use in the space–time analysis of data at scales that were inconceivable two decades ago (e.g., Zammit-Mangion et al., 2015).

Let $\boldsymbol{\eta}$ have a non-degenerate multivariate Gaussian distribution with precision matrix $\mathbf{Q}$, and encode the pairwise conditional dependence properties of $\boldsymbol{\eta}$ in the form of a graph $\mathcal{G}_Q = \{\mathcal{V}, \mathcal{E}_Q\}$, where $\mathcal{V}$ indexes the elements of $\boldsymbol{\eta}$, and $(i, j) \notin \mathcal{E}_Q$ exactly when $\eta_i \perp \eta_j \mid \{\eta_k : k \neq i, j\}$, for $i \neq j$. As is well-known (see, e.g., Rue and Held, 2005, Theorem 2.2), this graph defines the zeros in $\mathbf{Q}$, for which $Q_{ij} = 0$ if and only if $i \not\sim j$ in $\mathcal{G}_Q$, for $i \neq j$. As the distribution of $\boldsymbol{\eta}$ is non-degenerate, the positivity condition of the Hammersley–Clifford Theorem holds (Besag, 1974), and $\mathcal{G}_Q$ also encodes the local Markov property and the global Markov property of $\boldsymbol{\eta}$.

☆ Reproducible code available as Supplementary Material (see Appendix A).

* Corresponding author.
E-mail addresses: azm@uow.edu.au (A. Zammit-Mangion), j.c.rougier@bristol.ac.uk (J. Rougier).

https://doi.org/10.1016/j.csda.2018.02.001
0167-9473/

In this article we consider the case when the GMRF is used to encode prior belief on the quantity $\boldsymbol{\eta}$ through $\mathbf{Q}$, that may itself be a function of a small number of parameters that need to be estimated. We further assume that $\boldsymbol{\eta}$ is not directly observed; instead, a linear combination of $\boldsymbol{\eta}$, $\mathbf{B}\boldsymbol{\eta}$, is observed in the presence of noise. Denote the data vector as $\mathbf{Z}$. The two-level hierarchical model we consider is

$$\mathbf{Z} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\epsilon},$$
$$\boldsymbol{\eta} \sim \text{Gau}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Q}^{-1}),$$

where $\mathbf{X}$ are covariates, $\boldsymbol{\beta}$ are regression coefficients, and $\boldsymbol{\epsilon}$ is Gaussian, uncorrelated, measurement error with diagonal precision matrix $\mathbf{R}$.

It is an immediate result that if $\mathbf{X}$, $\mathbf{B}$, $\mathbf{Q}$, $\mathbf{R}$ and $\boldsymbol{\beta}$ are known, then the precision matrix of $\boldsymbol{\eta} \mid \mathbf{Z}$ is $\mathbf{P} := \mathbf{B}^T\mathbf{R}\mathbf{B} + \mathbf{Q}$. It is also well known that $\text{var}(\eta_i \mid \mathbf{Z})$ can be easily found from the sparse Cholesky factor of $\mathbf{P}$ using the Takahashi equations, without computing $\mathbf{S} := \mathbf{P}^{-1}$ directly (Takahashi et al., 1973; Erisman and Tinney, 1975; Rue and Martino, 2007). Frequently, however, we wish to compute prediction variances of *linear combinations* of $\boldsymbol{\eta}$, for example over sub-groups of variables, or over regions in a spatial domain in what is sometimes referred to as the change of support problem (Wikle and Berliner, 2005). This computation is always needed in the ubiquitous case when the spatial field is modelled as a sum of basis functions, and where a GMRF prior is placed on the basis-function coefficients.

This article investigates the use of sparse linear algebra methods for the computation of the marginal variances of $\mathbf{A}\boldsymbol{\eta} \mid \mathbf{Z}$, that is, $\mathbf{d} := \text{diag}(\mathbf{A}\mathbf{S}\mathbf{A}^T)$, when $\mathbf{A}$ is nonnegative and when $\mathbf{P}$ is such that its Cholesky factor can be computed. Specifically, it establishes the conditions on $\mathbf{A}$ under which $\mathbf{d} = \text{diag}(\mathbf{A}\tilde{\mathbf{S}}\mathbf{A}^T)$, where $\tilde{\mathbf{S}}$ is a sparse subset of $\mathbf{S}$ containing a sparsity pattern that is in general identical to that of the Cholesky factor of $\mathbf{P}$, also a by-product of the Takahashi equations. We find that in several situations of practical importance, this computation simplification facilitates the evaluation of conditional variances over linear combinations where direct computation is only possible in a massively parallel computing environment, and where conditional simulation, while feasible, is inaccurate when the number of simulations is limited to a reasonable value.

Sparse inverse subsets are frequently used to facilitate computation in estimation frameworks (e.g., Gilmour et al., 1995; Kiiveri and De Hoog, 2012; Cseke et al., 2016). They are particularly useful for computing trace operations appearing in estimating equations of the form $\mathbf{d}^T\mathbf{1} = \text{tr}(\mathbf{S}\mathbf{A}^T\mathbf{A})$. Bolin et al. (2009) noted that if $\mathbf{A} = \mathbf{B}$ then $\tilde{\mathbf{S}}$ necessarily contains the required elements to compute the trace, and thus replaced $\mathbf{S}$ with $\tilde{\mathbf{S}}$ when computing this trace operation in the M-step of an expectation–maximisation algorithm. Vanhatalo et al. (2010) solved the related problem of computing $\text{tr}(\mathbf{S}\mathbf{D})$ where $\mathbf{D}$ has the same sparsity pattern as $\mathbf{S}^{-1}$, by replacing $\mathbf{S}$ with $\tilde{\mathbf{S}}$. In a similar vein, Grigorievskiy et al. (2016) computed the block-diagonal inverse subset of $\mathbf{S}$ to find the trace when both $\mathbf{S}^{-1}$ and $\mathbf{D}$ are block tridiagonal. In this article we instead focus on the computation of all of $\mathbf{d}$, which in a spatial context are the prediction error variances at different levels of spatial aggregation (as determined by $\mathbf{A}$).

Our main result is presented in Section 2 while a complexity analysis is given in Section 3. In Section 4 we then develop the framework required for applying this result in a spatial-analysis setting, and demonstrate its use in several case studies. These studies consider conditional-autoregressive models, `LatticeKrig` models, statistical downscaling, and spatial-random effects models. Section 5 concludes with a brief mention of other approaches currently being investigated for when the sparse Cholesky factor is too large to compute.

2. Main result

Let $\mathbf{A}$ and $\mathbf{B}$ be nonnegative matrices, and let $\mathbf{Q}$ and $\mathbf{R}$ be positive definite symmetric matrices, where the dimensions of all matrices are implicit in what follows. Define $\mathbf{P} := \mathbf{B}^T\mathbf{R}\mathbf{B} + \mathbf{Q}$; hence $\mathbf{P}$ is positive definite even if $\mathbf{B}$ is not full rank. Further, define $\mathbf{S} := \mathbf{P}^{-1}$. Our objective is to compute the vector $\mathbf{d} := \text{diag}(\mathbf{A}\mathbf{S}\mathbf{A}^T)$. To summarise,

$$\mathbf{P} := \mathbf{B}^T\mathbf{R}\mathbf{B} + \mathbf{Q}, \tag{1a}$$
$$\mathbf{S} := \mathbf{P}^{-1}, \tag{1b}$$
$$\mathbf{d} := \text{diag}(\mathbf{A}\mathbf{S}\mathbf{A}^T). \tag{1c}$$

This section presents a theorem relating $\mathbf{d}$ to the sparsity structure of $\mathbf{P}$ and $\mathbf{A}$.

The following simple lemma establishes a necessary and sufficient condition for $\mathbf{d}$ to be invariant to any specified element of $\mathbf{S}$, in terms of the elements of $\mathbf{A}$.

Lemma 2.1. *The vector* $\mathbf{d}$ *is invariant to* S_{jk} *if and only if* $[\mathbf{A}^T\mathbf{A}]_{jk} = 0$.

Proof. The ith element of $\mathbf{d}$ is

$$d_i = \sum_j \sum_k A_{ij} S_{jk} A_{ik} = \sum_j \sum_k (A_{ij} A_{ik}) S_{jk}.$$

Hence d_i is invariant to S_{jk} if and only if $A_{ij}A_{ik} = 0$. Therefore the entire vector $\mathbf{d}$ is invariant to S_{jk} if and only if $A_{ij}A_{ik} = 0$ for all i, or, because $\mathbf{A}$ is nonnegative, $[\mathbf{A}^T\mathbf{A}]_{jk} = 0$. □