NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

13-18

A Diagonally Weighted Matrix Norm between Two Covariance Matrices

Noel Cressie and Cécile Hardouin

Copyright © 2018 by the National Institute for Applied Statistics Research Australia, UOW. Work in progress, no part of this paper may be reproduced without permission from the Institute.

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522, Australia Phone +61 2 4221 5076, Fax +61 2 4221 4998. Email: karink@uow.edu.au

A Diagonally Weighted Matrix Norm Between Two Covariance Matrices

Noel Cressie[a], Cécile Hardouin[b]

[a]*NIASRA, University of Wollongong, Australia*
[b]*MODAL'X, Université Paris Nanterre, France*

Abstract

The square of the Frobenius norm of a matrix A is defined as the sum of squares of all the elements of A. An important application of the norm in statistics is when A is the difference between a target (estimated or given) covariance matrix and a parameterized covariance matrix, whose parameters are chosen to minimize the Frobenius norm. In this article, we investigate weighting the Frobenius norm by putting more weight on the diagonal elements of A, with an application to spatial statistics. We find the spatial random effects (SRE) model that is closest, according to the the weighted Frobenius norm between covariance matrices, to a particular stationary Matérn covariance model.

Keywords: condition number, Fixed Rank Kriging, Frobenius norm, Q-R decomposition, spatial random effects model

1. Introduction

Fundamental to all of statistics is the modeling of a mean vector and a covariance matrix. This article is concerned with how close two covariance matrices are to each other, for the purposes of model calibration or parameter estimation. In particular, we consider the Frobenius norm and develop a new, weighted version of it that puts more weight on the diagonal elements, hence giving more emphasis to variances than covariances.

Spatial statistics has become important in many applications, particularly in the earth and environmental sciences. Better sensors, for example on satellites, have led to a rapid increase in the size n of spatial data sets. Kriging (Matheron, 1962) is an optimal method of spatial prediction that filters out noise and fills in gaps in the data, but the kriging equations involve the inverse of the $n \times n$ data covariance matrix $\mathbf{\Sigma}$. In general, the computations to obtain the kriging predictor and kriging variance are not scalable, usually of $O(n^3)$. Solutions to this problem include reduced-dimension methods (see Wikle, 2010, for a review) and the use of sparse precision matrices (Lindgren et al., 2011; Nychka et al., 2015). One of the reduced-dimension methods is based on the spatial random effects (SRE) model, which is a spatial process given by a random linear combination of r known spatial basis functions, where r is fixed and relatively small (Cressie and Johannesson, 2006, 2008). The resulting spatial prediction, called Fixed Rank Kriging (FRK), has a computational complexity of just $O(nr^2) = O(n)$, for r fixed.

Email addresses: ncressie@uow.edu.au (Noel Cressie), hardouin@parisnanterre.fr (Cécile Hardouin)

Preprint submitted to Spatial statistics *October 29, 2018*

The SRE class of spatial covariance matrices is chosen to illustrate the methodology presented in this article. One way to estimate the SRE-model parameters is via an EM algorithm, which requires parametric (usually Gaussian) assumptions. Alternatively, the SRE-model parameters can be estimated via minimizing a Frobenius matrix norm (Cressie and Johannesson, 2008) which, in this article, we generalize to a diagonally weighted Frobenius norm.

In Section 2, we present the Frobenius norm (F-norm) and its use for estimating covariance parameters; then we define a diagonally weighted version, the D-norm. Section 3 reviews briefly the spatial random effects (SRE) model and recalls the least-F-norm estimate of its parameters. In Section 4, we derive new estimating equations for the least-D-norm estimate of the SRE model's parameters, for which we obtain an analytic solution for estimating the covariance matrix of the random effects. Section 5 presents a study that investigates the effects of the extra weight added to the diagonal, and we obtain least-F-norm and least-D-norm fits of the covariance matrix of the random effects. Then we compare the two fitted spatial covariance matrices by computing Kullback-Leibler divergences from the given true Gaussian distribution. We also compare various matrix norms of the difference between the true spatial covariance matrix and the fitted spatial covariance matrix, as well as the condition numbers of the two fitted SRE-parameter covariance matrices. We finally give heuristics to choose the diagonal weights depending on the strength of the spatial dependence. The paper ends with a discussion in Section 6.

2. The Frobenius norm and its diagonally weighted version

2.1. The Frobenius norm (F-norm)

Let $\text{tr}(\mathbf{A})$ denote the trace operator that sums the diagonal elements of a square matrix $\mathbf{A}$. The Frobenius norm (F-norm) of an $n \times n$ matrix $\mathbf{A}$ is defined as,

$$\|\mathbf{A}\|_F \equiv \left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2 \right)^{1/2} = (\text{tr}(\mathbf{A}'\mathbf{A}))^{1/2} . \tag{1}$$

Notice that each element of $\mathbf{A}$ is weighted exactly the same. One way to introduce non-negative weights $\{w_1, \ldots, w_n\}$ is to take the F-norm of $\mathbf{WAW}$ or of $\mathbf{WA}$, where $\mathbf{W}$ is a diagonal matrix with $\{w_1^{1/2}, \ldots, w_n^{1/2}\}$ down the diagonal. For each of these options, it is not possible to put extra emphasis on the diagonal elements of $\mathbf{A}$. In this article, we propose a way to do this and call it the Diagonally Weighted Frobenius norm, that we shall denote D-norm, short for DWF-norm.

Now, suppose we wish to fit $\boldsymbol{\theta}$ by minimizing the norm of the difference, $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}(\boldsymbol{\theta})$, where $\boldsymbol{\Sigma}_0$ is a target covariance matrix and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is a covariance matrix depending on unknown parameters $\boldsymbol{\theta}$. In the application given in Section 5, $\boldsymbol{\Sigma}_{0,ij} = C(\mathbf{s}_i, \mathbf{s}_j)$ where C is a given covariance function. In other settings, if $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ is an n-dimensional spatial process, then suppose we model $\text{cov}(\mathbf{Z}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$; if $\mathbf{Z}$ is observed independently m times, resulting in data $\mathbf{Z}_1, ..., \mathbf{Z}_m$, then we could choose for $\boldsymbol{\Sigma}_0$ the non-parametric estimator,

$$\boldsymbol{\Sigma}_0 = \hat{\boldsymbol{\Sigma}}_m \equiv (1/m) \sum_{k=1}^{m} \left(\mathbf{Z}_k - \bar{\mathbf{Z}}\right)\left(\mathbf{Z}_k - \bar{\mathbf{Z}}\right)' , \tag{2}$$

where $\bar{\mathbf{Z}}$ is the empirical mean, $\bar{\mathbf{Z}} \equiv \sum_{k=1}^{m} \mathbf{Z}_k / m$. For example, Sampson and Guttorp (1992) use replicates $\{\mathbf{Z}_t : t = 1 \ldots m\}$ (over time) to obtain $\boldsymbol{\Sigma}_0$ given by (2).

Suppose that the target covariance matrix $\mathbf{\Sigma}_0$ is obtained from the data, for example $\hat{\mathbf{\Sigma}}_m$ in (2). A least-F-norm estimator of covariance parameters, $\boldsymbol{\theta}$, is defined as:

$$\hat{\boldsymbol{\theta}} \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} ||\mathbf{\Sigma}_0 - \mathbf{\Sigma}(\boldsymbol{\theta})||_F^2 \,, \tag{3}$$

where Θ is the parameter space of $\boldsymbol{\theta}$. This is a semiparametric alternative to finding a maximum likelihood estimator of $\boldsymbol{\theta}$ or a restricted maximum likelihood estimator of $\boldsymbol{\theta}$, where typically a parametric assumption is made that data are distributed as a multivariate Gaussian distribution. If (2) is used in (3), the only distributional assumption required is the existence of the first two moments of the elements $\{Z_i : i = 1, \ldots, n\}$ of $\mathbf{Z}$.

We shall now separate the variances from the covariances. Define

$$\mathbf{V}(\boldsymbol{\theta}_v) \equiv \text{diag}(\mathbf{\Sigma}(\boldsymbol{\theta})) \,,$$

where diag($\mathbf{B}$) is a diagonal matrix with $\{(\mathbf{B})_{ii} : i = 1, \ldots, n\}$ down the diagonal, and $\boldsymbol{\theta}_v \in \Theta_v \subset \Theta$ are parameters of $(var(Z_1), \ldots, var(Z_n))'$. Then, when the target covariance matrix $\mathbf{\Sigma}_0$ is obtained from the data, a least-F-norm estimator, $\hat{\boldsymbol{\theta}}_v$, can be obtained by minimizing with respect to $\boldsymbol{\theta}_v$,

$$||\text{diag}(\mathbf{\Sigma}_0) - \mathbf{V}(\boldsymbol{\theta}_v)||_F^2 = \text{tr}\left(\text{diag}\left(\mathbf{\Sigma}_0 - \mathbf{\Sigma}(\boldsymbol{\theta})\right)' \,\text{diag}\left(\mathbf{\Sigma}_0 - \mathbf{\Sigma}(\boldsymbol{\theta})\right)\right) \,.$$

That is,

$$\hat{\boldsymbol{\theta}}_v = \arg\min_{\boldsymbol{\theta}_v\in\Theta_v} ||\text{diag}(\mathbf{\Sigma}_0) - \mathbf{V}(\boldsymbol{\theta}_v)||_F^2 \,. \tag{4}$$

2.2. A diagonally weighted Frobenius norm (D-norm)

Motivated by (3) and (4), we introduce a diagonally weighted Frobenius norm (D-norm), $||\mathbf{A}||_D$, through

$$||\mathbf{A}||_D^2 \equiv \text{tr}(\mathbf{A}'\mathbf{A}) + \lambda^2 \text{tr}\left(\text{diag}(\mathbf{A})'\text{diag}(\mathbf{A})\right) = ||\mathbf{A}||_F^2 + \lambda^2 ||\text{diag}(\mathbf{A})||_F^2 \,, \tag{5}$$

where λ^2 is fixed and, hence, the D-norm depends on it. Note that it is straightforward to show that $||.||_D$ defined by (5) satisfies all the properties of a norm. Consequently, for $\lambda^2 > 0$, $||\mathbf{\Sigma}_0 - \mathbf{\Sigma}(\boldsymbol{\theta})||_D^2$ puts more emphasis on matching the variances than the covariances. Once again, suppose that the target covariance matrix $\mathbf{\Sigma}_0$ is obtained from the data. Then define the least-D-norm estimator of $\boldsymbol{\theta}$ as follows:

$$\hat{\boldsymbol{\theta}}(\lambda^2) \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} ||\mathbf{\Sigma}_0 - \mathbf{\Sigma}(\boldsymbol{\theta})||_D^2 \,, \tag{6}$$

where $\hat{\boldsymbol{\theta}}(0)$ is given by (3), and $\hat{\boldsymbol{\theta}}(\infty)$ is given by (4). In general, the estimator $\hat{\boldsymbol{\theta}}(\lambda^2)$ depends on λ^2, namely the amount of extra weight put on the diagonal elements.

3. Minimizing the F-norm to estimate parameters of the SRE model

We first define the spatial random effects (SRE) model and fit or estimate its covariance parameters by minimizing the Frobenius norm (F-norm).

3.1. The SRE model

Suppose that $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ are spatial data on a finite set of locations, $D \equiv \{\mathbf{s}_i : i = 1, \ldots, n\} \subset \mathbb{R}^d$, in a d-dimensional Euclidean space. We write $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))'$, where now Z_i defined in Section 2 has an explicit spatial index $\mathbf{s}_i$; that is, $Z_i \equiv Z(\mathbf{s}_i)$, for $i = 1, \ldots, n$. We posit the following decomposition for $Z(\cdot)$: For $\mathbf{s} \in D$,

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \varepsilon(\mathbf{s}) \,, \tag{7}$$

$$Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})'\boldsymbol{\beta} + W(\mathbf{s}) \,, \tag{8}$$

where $\mathbf{X}(\mathbf{s})'\boldsymbol{\beta}$ is the large-scale spatial variation due to p covariates, $\mathbf{X}(\cdot) \equiv (X_1(\cdot), ..., X_p(\cdot))'$, and the terms $\varepsilon(\cdot)$ and $W(\cdot)$ represent respectively the measurement error in (7) and the small-scale variation in (8). Here, both are assumed to have mean zero. We assume an SRE model for $W(\cdot)$, which is given by (Cressie and Johannesson, 2006, 2008):

$$W(\mathbf{s}) = \mathbf{S}(\mathbf{s})'\boldsymbol{\eta} + \xi(\mathbf{s}) \,;\ \mathbf{s} \in D \,,$$

where $\mathbf{S}(\cdot) \equiv (S_1(\cdot), ..., S_r(\cdot))'$ is a vector of pre-specified, known spatial basis functions; $\boldsymbol{\eta} \equiv (\eta_1, ..., \eta_r)'$ is a vector of random effects with mean zero and positive-definite covariance matrix $\mathbf{K}$, and $\xi(\cdot)$ represents the fine-scale variation in the process $Y(\cdot)$. It is assumed that $\xi(\cdot)$ has mean zero and correlation zero at distinct locations. That is, $\text{cov}(\xi(\mathbf{s}), \xi(\mathbf{u})) = \sigma^2_\xi V(\mathbf{s}) 1(\mathbf{s} = \mathbf{u})$, where $\sigma^2_\xi > 0$ is an unknown parameter, $V(\cdot) > 0$ is assumed known, and $1(\cdot)$ is an indicator function. Finally, $\xi(\cdot)$ is assumed to be statistically independent of $\boldsymbol{\eta}$.

In this article, our interest is in the $n \times n$ covariance matrix $\text{cov}\,((Z(\mathbf{s}) : \mathbf{s} \in D)') \equiv \boldsymbol{\Sigma}(\boldsymbol{\theta})$, where $Z(\cdot)$ is given by (7) and (8). Hence, we can assume that $\mathbf{X}(\cdot) \equiv \mathbf{0}$, since any fixed effect is ignored when calculating covariances. Then the model (8) reduces to

$$Z(\mathbf{s}) = \mathbf{S}(\mathbf{s})'\boldsymbol{\eta} + \xi(\mathbf{s}) + \varepsilon(\mathbf{s});\ \mathbf{s} \in D \,, \tag{9}$$

which in vector form can be written as

$$\mathbf{Z} = \mathbf{S}\boldsymbol{\eta} + \boldsymbol{\xi} + \boldsymbol{\varepsilon} \,, \tag{10}$$

where the three vectors on the right-hand side are mutually independent. In (10), $E(\boldsymbol{\eta}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\eta}) = \mathbf{K}$; $E(\boldsymbol{\xi}) = \mathbf{0}$, and $\text{cov}(\boldsymbol{\xi}) = \sigma^2_\xi \mathbf{V}$, where $\mathbf{V}$ is a known diagonal matrix with $V(\mathbf{s}_1), \ldots, V(\mathbf{s}_n)$ down the diagonal; and $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\varepsilon}) = \sigma^2_\varepsilon \mathbf{I}_n$, where $\mathbf{I}_n$ is the n-dimensional identity matrix. Hence,

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbf{SKS}' + \sigma^2_\xi \mathbf{V} + \sigma^2_\varepsilon \mathbf{I}_n \,, \tag{11}$$

where $\boldsymbol{\theta} = (\mathbf{K}, \sigma^2_\xi)$. There is often an identifiability problem with estimating σ^2_ξ and σ^2_ε, which is resolved by assuming σ^2_ε is known; we shall make that assumption here. In (11), parameters are $\boldsymbol{\theta} = (\mathbf{K}, \sigma^2_\xi) \in \Theta \equiv \{(\mathbf{K}, \sigma^2_\xi) :\ \mathbf{K} \text{ positive-definite, and } \sigma^2_\xi > 0\}$.

3.2. Fitting SRE covariance parameters using the F-norm

The covariance parameters in the SRE model are given by $\mathbf{K}$ and σ^2_ξ in (11). For a target covariance matrix $\boldsymbol{\Sigma}_0$, we wish to fit $\boldsymbol{\theta} = (\mathbf{K}, \sigma^2_\xi)$ by minimizing the norm of the difference, $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}(\boldsymbol{\theta})$. Without loss of generality, we simplify (11) by putting $\sigma^2_\varepsilon = 0$ and $\mathbf{V} = \mathbf{I}_n$. Otherwise, our results still hold, albeit with more complicated formulas. Hence, our goal is to find $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{K}}, \hat{\sigma}^2_\xi) \in \Theta$, by minimizing $\|\boldsymbol{\Sigma}_0 - \mathbf{SKS}' - \sigma^2_\xi \mathbf{I}_n\|_F$; the restriction to the parameter space Θ means that $\hat{\mathbf{K}}$ is positive-definite and $\hat{\sigma}^2_\xi > 0$. Write $\mathbf{S} = \mathbf{QR}$, the Q-R decomposition of $\mathbf{S}$ (i.e., $\mathbf{Q}$ is an $n \times r$ orthonormal matrix, and $\mathbf{R}$ is a non-singular $r \times r$ upper-triangular matrix), and define the vec operator $\text{vec}(\mathbf{B}) \equiv (\mathbf{b}'_1 \mathbf{b}'_2 \ldots \mathbf{b}'_n)'$ of the matrix $\mathbf{B} = (\mathbf{b}_1 \mathbf{b}_2 \ldots \mathbf{b}_n)$.

The following result gives analytic, closed-form expressions for $\hat{\mathbf{K}}$ and $\hat{\sigma}^2_\xi$.

Proposition 1. *Minimum F-norm estimator.*

Recall $(\hat{\mathbf{K}}, \hat{\sigma}^2_\xi) \equiv \arg\min_{\boldsymbol{\theta} \in \Theta} \|\boldsymbol{\Sigma}_0 - \mathbf{SKS}' - \sigma^2_\xi \mathbf{I}_n\|^2_F$. *Then*

$$\hat{\sigma}^2_\xi = \frac{(\text{vec}(\mathbf{QQ}'\boldsymbol{\Sigma}_0\mathbf{QQ}' - \boldsymbol{\Sigma}_0))' \,\text{vec}(\mathbf{QQ}' - \mathbf{I}_n)}{\|\mathbf{QQ}' - \mathbf{I}_n\|^2_F} \,, \tag{12}$$

and

$$\hat{\mathbf{K}} = \mathbf{R}^{-1}\mathbf{Q}'(\boldsymbol{\Sigma}_0 - \hat{\sigma}^2_\xi \mathbf{I}_n)\mathbf{Q}(\mathbf{R}^{-1})' \,, \tag{13}$$

provided $\boldsymbol{\Sigma}_0 - \hat{\sigma}^2_\xi \mathbf{I}_n$ *is positive-definite and the right-hand side of* (12) *is positive.*

The proof is given in the Appendix. In practice, the first condition can be checked by verifying positive-definiteness of the $r \times r$ matrix on the right-hand side of (13).

4. Fitting SRE covariance parameters using the D-norm

From (5),

$$\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}(\boldsymbol{\theta})\|^2_D = \|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}(\boldsymbol{\theta})\|^2_F + \lambda^2 \|\text{diag}(\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}(\boldsymbol{\theta}))\|^2_F \,, \tag{14}$$

where recall from (11) that $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbf{SKS}' + \sigma^2_\xi \mathbf{I}$, for $\boldsymbol{\theta} = (\mathbf{K}, \sigma^2_\xi)$, $\mathbf{K}$ positive-definite, and $\sigma^2_\xi > 0$. For λ^2 given, a least-D-norm estimate of $\boldsymbol{\theta}$ is the parameter value that minimizes (14) above.

Let us write $\mathbf{Q}' \equiv (\mathbf{Q}_1 \ldots \mathbf{Q}_n)$, and let $\mathbf{u}$ be an n-dimensional vector. We define

$$\mathbf{g}(\mathbf{Q}) \equiv (\text{vec}(\mathbf{Q}_1\mathbf{Q}'_1), \ldots, \text{vec}(\mathbf{Q}_n\mathbf{Q}'_n)) \begin{pmatrix} \text{vec}(\mathbf{Q}_1\mathbf{Q}'_1)' \\ \vdots \\ \text{vec}(\mathbf{Q}_n\mathbf{Q}'_n)' \end{pmatrix} \tag{15}$$

and

$$\mathbf{h}(\mathbf{Q}, \mathbf{u}) \equiv (\text{vec}(\mathbf{Q}_1\mathbf{Q}'_1), \ldots, \text{vec}(\mathbf{Q}_n\mathbf{Q}'_n))\, \mathbf{u} \,. \tag{16}$$

The matrix $\mathbf{g}$ defined in (15) is $r^2 \times r^2$, and $\mathbf{h}(\mathbf{Q}, \mathbf{u})$ defined in (16) is an r^2-dimensional vector.

Now, let us define the $r \times r$ matrix $\hat{\mathbf{K}}^*$ through the vec operator:

$$\text{vec}(\hat{\mathbf{K}}^*(\sigma^2_\xi; \lambda^2)) \equiv (\mathbf{I}_{r^2} + \lambda^2 \mathbf{g}(\mathbf{Q}))^{-1} \left\{ \text{vec}(\mathbf{Q}'(\boldsymbol{\Sigma}_0 - \sigma^2_\xi \mathbf{I}_n)\mathbf{Q}) + \lambda^2 \mathbf{h}(\mathbf{Q}, \text{diag}(\boldsymbol{\Sigma}_0 - \sigma^2_\xi \mathbf{I}_n)) \right\} \,, \tag{17}$$

and hence define

$$\hat{\mathbf{K}}(\sigma^2_\xi; \lambda^2) \equiv \mathbf{R}^{-1}\hat{\mathbf{K}}^*(\sigma^2_\xi; \lambda^2)(\mathbf{R}^{-1})' \,. \tag{18}$$

The following result gives analytic, closed-form expressions for $\hat{\mathbf{K}}(\lambda^2)$ and $\hat{\sigma}^2_\xi(\lambda^2)$, for a given λ^2. The proof is given in the Appendix.

Proposition 2. *Minimum D-norm estimator.*

For a given λ^2, $\hat{\boldsymbol{\theta}}(\lambda^2) \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{\Sigma}_0 - \mathbf{SKS}' - \sigma^2_\xi \mathbf{I}_n\|^2_D$ *is given by*

$$\hat{\sigma}^2_\xi(\lambda^2) \ = \arg\min_{\sigma^2_\xi>0} \|\boldsymbol{\Sigma}_0 - \mathbf{S}\hat{\mathbf{K}}(\sigma^2_\xi;\lambda^2)\mathbf{S}' - \sigma^2_\xi \mathbf{I}_n\|^2_D \ , \tag{19}$$

and

$$\hat{\mathbf{K}}(\lambda^2) \ = \hat{\mathbf{K}}(\hat{\sigma}^2_\xi(\lambda^2);\lambda^2)\ , \tag{20}$$

provided $\boldsymbol{\Sigma}_0 - \hat{\sigma}^2_\xi(\lambda^2)\mathbf{I}_n$ *is positive-definite.*

Importantly, the minimization in (19) is restricted to those $\sigma^2_\xi > 0$ that yield a positive-definite $\hat{\mathbf{K}}(\sigma^2_\xi;\lambda^2)$. From (17), this is guaranteed by considering only those $\sigma^2_\xi > 0$ such that $\boldsymbol{\Sigma}_0 - \sigma^2_\xi\mathbf{I}_n$ is positive-definite, which is the same condition given in Section 3.2 for the minimum F-norm estimator. Because of the closed-form expression for $\hat{\mathbf{K}}(\sigma^2_\xi;\lambda^2)$, the minimization in (19) is only with respect to the one-dimensional parameter $\sigma^2_\xi > 0$, and it can be easily obtained by a golden search for example.

5. Application

In this section, we illustrate the advantage of using the D-norm in fitting an SRE model (9) to the well known exponential-covariance model, which is a particular case of the Matérn covariance model. We consider a two-dimensional lattice $D = \{\mathbf{s}_{ij} : i, j = 1, \dots, N\}$ with $N = 100$; that is, $n = 10^4$. We choose bisquare functions for the spatial basis functions, with three resolutions, the centers being regularly spaced within a resolution. The generic expression for these basis functions is,

$$S_{j(l)}(\mathbf{s}) = \begin{cases} 1 - \dfrac{\|\mathbf{s}-\mathbf{c}_{j(l)}\|}{r_l} & \text{if } \|\mathbf{s}-\mathbf{c}_{j(l)}\| \le r_l \\ 0 & \text{otherwise,} \end{cases}$$

where $c_{j(l)}$ is the jth centre point of the lth resolution, for $l = 1, 2, 3$, and $\|\mathbf{s} - \mathbf{u}\|$ is the Euclidean distance between two locations $\mathbf{s}$ and $\mathbf{u}$. The number of basis functions used at the three resolutions are, respectively 5, 16, and 49. Consequently, the dimension of the reduced space is $r = 70$. The radius r_l of the lth resolution bisquare function equals 1.5 times the shortest distance between center points of this resolution, allowing overlap between the basis functions.

We want to find σ^2_ξ and $\mathbf{K}$ that minimize the norm of the difference, $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}(\sigma^2_\xi, \mathbf{K})$, where the target covariance $\boldsymbol{\Sigma}_{0,ij} = C(\mathbf{s}_i, \mathbf{s}_j)$ is obtained from an exponential covariance function to which we choose to add a nugget effect. That is,

$$C(\mathbf{u}, \mathbf{v}) = c \exp\left(-\frac{\| \mathbf{u} - \mathbf{v} \|}{\varphi}\right) + a\ , \tag{21}$$

where c is the sill, φ is the scale parameter, and $a \ge 0$ is the nugget effect. Here we specify $c = 1$ (without loss of generality), and φ ranges from 5 to 70, to capture weak to strong spatial dependence, respectively. We adopt this strategy because the spatial dependence in the exponential covariance function given by (21) is well understood. Our goal here is not parameter estimation, but it is to find σ^2_ξ and $\mathbf{K}$ that approximate the given covariance model $\boldsymbol{\Sigma}_0$ with the "nearest" SRE covariance model.

We obtain $\hat{\mathbf{K}}_F$ and $\hat{\sigma}^2_{\xi,F}$ defined in (12) and (13), by minimizing $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}(\sigma^2_\xi, \mathbf{K})\|_F$; and we obtain $\hat{\mathbf{K}}_D(\lambda^2)$ and $\hat{\sigma}^2_{\xi,D}(\lambda^2)$ defined in (19) and (20), by minimizing $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}(\sigma^2_\xi, \mathbf{K})\|_D$, for various choices of λ^2.

To compare the accuracy of the fits obtained from using the F-norm and the D-norm, we use a number of measures. Recall the Kullback-Leibler divergence, $D_{KL}(P_0|Q)$, where P_0 is a Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_0$, and Q is a Gaussian distribution of the same dimension with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_Q$, as follows:

$$D_{KL}(P_0|Q) \equiv -\frac{1}{2}\log\left(\frac{\det \boldsymbol{\Sigma}_0}{\det \boldsymbol{\Sigma}_Q}\right) - \frac{n}{2} + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_0\right) . \tag{22}$$

In our use of the Kullback-Leibler divergence in (22), $\boldsymbol{\Sigma}_Q$ is one or other of

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_F) = \mathbf{S}\hat{\mathbf{K}}_F\mathbf{S}' + \hat{\sigma}^2_{\xi,F}\mathbf{I}_n , \tag{23}$$

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_D(\lambda^2)) = \mathbf{S}\hat{\mathbf{K}}_D(\lambda^2)\mathbf{S}' + \hat{\sigma}^2_{\xi,D}(\lambda^2)\mathbf{I}_n . \tag{24}$$

One way that the efficacy of the D-norm fit can be compared to the F-norm fit is through the relative Kullback-Leibler divergence,

$$E_{KL} \equiv \frac{D_{KL}(P_0|Q(\hat{\boldsymbol{\theta}}_F))}{D_{KL}(P_0|Q(\hat{\boldsymbol{\theta}}_D(\lambda^2)))} . \tag{25}$$

Another way is through relative matrix norms. For example, define

$$E_2 \equiv \frac{\| \boldsymbol{\Sigma}_0 - \hat{\boldsymbol{\Sigma}}_F \|_2}{\| \boldsymbol{\Sigma}_0 - \hat{\boldsymbol{\Sigma}}_D(\lambda^2) \|_2} , \tag{26}$$

and

$$E_{\max} \equiv \frac{\| \boldsymbol{\Sigma}_0 - \hat{\boldsymbol{\Sigma}}_F \|_{\max}}{\| \boldsymbol{\Sigma}_0 - \hat{\boldsymbol{\Sigma}}_D(\lambda^2) \|_{\max}} , \tag{27}$$

where $\| \mathbf{A} \|_{\max} \equiv \max_{i,j} |a_{ij}|$ and $\| \mathbf{A} \|_2 \equiv \sigma_{\max}(\mathbf{A})$, the largest singular value of the matrix $\mathbf{A}$. The following inequality holds between the norms we consider:

$$\| \cdot \|_{\max} \leq \| \cdot \|_2 \leq \| \cdot \|_F \leq \| \cdot \|_D . \tag{28}$$

Another way to compare the D-norm to the F-norm is to examine the condition number of the fitted SRE covariance parameter $\hat{\mathbf{K}}$; define the relative condition number,

$$E_C \equiv \frac{\mathrm{cond}(\hat{\mathbf{K}}_F)}{\mathrm{cond}(\hat{\mathbf{K}}_D(\lambda^2))} , \tag{29}$$

where cond($\mathbf{A}$) is the 2-norm condition number of a matrix $\mathbf{A}$ (the ratio of the largest singular value of $\mathbf{A}$ to the smallest). A large condition number indicates a nearly singular matrix.

Our study that compares minimum D-norm fits to minimum F-norm fits is not a simulation; rather we computed the ratios E_{KL}, E_2, E_{max}, and E_C defined in (25), (26), (27), and (29), respectively, for various values of the factors φ, a, and λ^2 in a factorial design. The nugget effect a is defined in terms of proportion of the total variance; that is, $a = c\dfrac{p}{1-p}$, where $c = 1$ here and

$p \in \{0,\ 1/10,\ 1/3,\ 1/2,\ 2/3,\ 9/10\}$. The scale parameter $\varphi \in \Phi \equiv \{5, 10, 20, 30, 40, 50, 60, 70\}$; as φ increases from 5 to 70, it induces weak to strong spatial dependence. Finally, for the weights on the diagonal for the D-norm, we used smaller weights, $\lambda^2 \in \Lambda_1 \equiv \{0.1,\ 10,\ 20,\ 30, \ldots, 100\}$, , and larger weights, $\lambda^2 \in \Lambda_2 \equiv \{100k :\ k = 1, 2, \ldots, 10\}$.

We now summarize the results obtained. First, the nugget effect does not impact the values of the ratios $E_{max},\ E_2,\ E_C$, and only very slightly those of E_{KL}. Hence, we choose to present the following results with $a = 0$, and we have chosen to compare results here for scale parameter $\varphi \in \{5, 20, 40, 70\}$. Plots of E_{KL} and E_C against λ^2 are presented in Figure 1 and Figure 2; and plots of E_2 and E_{max} against λ^2 are presented in Figures 3 and 4. Figures 1 and 3 show the case $\lambda^2 \in \Lambda_1$, while Figures 2 and 4 show the case $\lambda^2 \in \Lambda_2$.

When limiting the comparison to how well the original covariance matrix $\mathbf{\Sigma}_0$ is fitted, it is clear that the D-norm performs in a similar manner to the F-norm, since E_{KL} and E_2 remain very close to 1. We have $0.9598 \leq E_{KL} \leq 1$. The smallest value of E_{KL} is obtained for $p = 90\%, \varphi = 70$, and $\lambda^2 = 1000$, but we have $E_{KL} \geq 0.984$ for $p \leq 80\%$, regardless of the values of φ and λ^2. Similarly, we always have $0.9924 \leq E_2 \leq 1.0015$.

Now, we highlight the advantage of the D-norm with respect to the max norm, $\|\ \|_{max}$, and the condition number of the matrix $\hat{\mathbf{K}}$. The ratios of E_{max} increase with φ and with λ^2. The values of E_{max} vary from 0.998 to 1.774; we have $E_{max} \geq 1.2$ for $\varphi \geq 40$ and $\lambda^2 \geq 100$, or $\varphi \geq 30$ and $\lambda^2 \geq 700$. So, globally we can say that the D-norm performs better than the F-norm with respect to the matrix norm $\|\ \|_{max}$. Let us now consider the values of E_C, which is defined in terms of the SRE model's covariance-matrix parameter. As before, the ratios of E_C increase with φ and λ^2; E_C increases from 0.9955 to 1.0621 for $\lambda^2 \in \Lambda_1$, and we achieve a gain of 30% for $\lambda^2 = 1000$, which is quite important. Also, the ratio E_C increases with φ; for instance, for $\lambda^2 = 500$, E_C increases from 0.9966 to 1.1985 for $\varphi \in \Phi$ and, for $\lambda^2 = 1000$, E_C increases from 1.0064 to 1.2967 for $\varphi \in \Phi$. While the D-norm condition number does not improve for weak spatial dependence, it becomes more and more efficient to use the D-norm as the spatial dependence strengthens.

We also conducted the same study, but with four resolutions, and a total of $r = 78$ basis functions, and we obtained similar results. We conclude that when the spatial dependence is moderate to strong, the D-norm should be used to fit the covariance parameters $\mathbf{K}$ and σ^2_ξ of an SRE model.

Finally, we present an empirical way of choosing λ^2 in Figure 5, where we plot $\lambda^2/\sqrt{n}$ against $\varphi/\sqrt{n}$ for different fixed ranges of E_C. We choose the relative condition number E_C, because the inverse of the matrix $\mathbf{K}$ is directly involved in the kriging equations, and hence, it is important that $\mathbf{K}$ not be ill-conditioned. We considered four ranges of values of E_C in Figure 5, namely $0.9 < E_C < 1.12$, $1.13 < E_C < 1.16$, $1.18 < E_C < 1.22$, and $E_C > 1.25$, resulting in "gains" of about 10 percent, 15 percent, 20 percent, and more than 25 percent, respectively. For each fixed range, we recorded for each value of $\varphi/\sqrt{n}$ the values of $\lambda^2/\sqrt{n}$ ensuring that E_C belongs to that range. Our main observation is that we need large values of λ^2 when the spatial dependence is moderate, and we need smaller values of λ^2 when the spatial dependence is strong. While no expression is derived linking $E_C,\ \lambda^2,\ \varphi$, and n, it can be seen that $\lambda^2/\sqrt{n} \geq (\varphi/\sqrt{n})^{-2}$ ensures that $E_C \geq 1.1$.

6. Discussion

Fitting covariance parameters of the SRE model can be achieved by using the Frobenius matrix norm (F-norm). This paper presents a diagonally weighted Frobenius matrix norm (D-

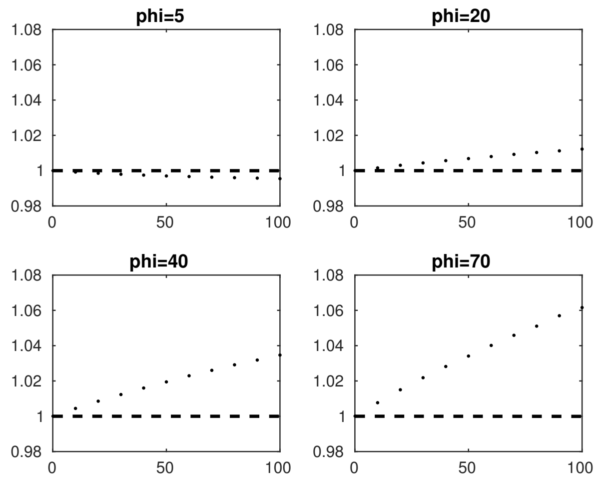

Figure 1: Plots of E_{KL} (–) and E_C (.) against $\lambda^2 \in \Lambda_1$ on the horizontal axis, for four values of $\varphi \in \Phi$.

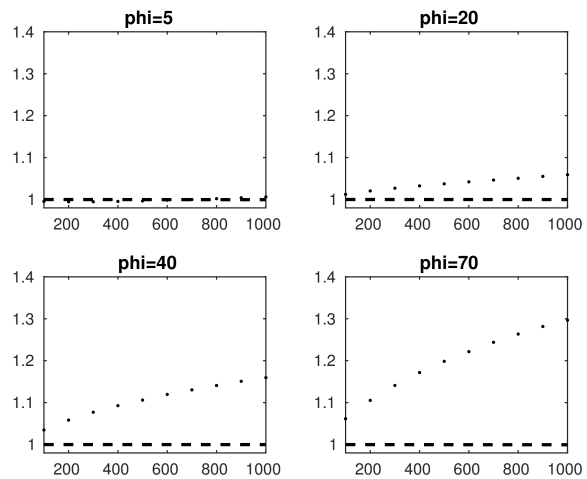

Figure 2: Plots of E_{KL} (–) and E_C (.) against $\lambda^2 \in \Lambda_2$ on the horizontal axis, for four values of $\varphi \in \Phi$.

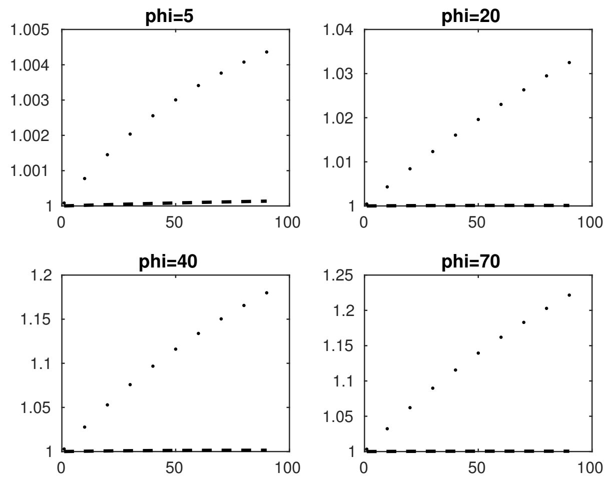

Figure 3: Plots of E_2 (–) and E_{max} (.) against $\lambda^2 \in \Lambda_1$ on the horizontal axis, for four values of $\varphi \in \Phi$.

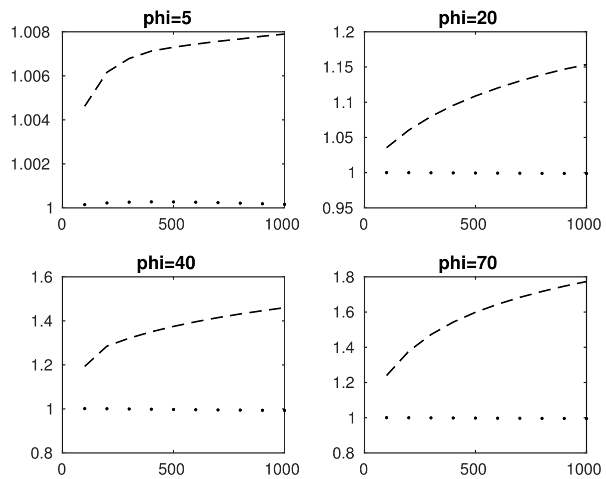

Figure 4: Plots of E_2 (–) and E_{max} (.) against $\lambda^2 \in \Lambda_2$ on the horizontal axis, for four values of $\varphi \in \Phi$.

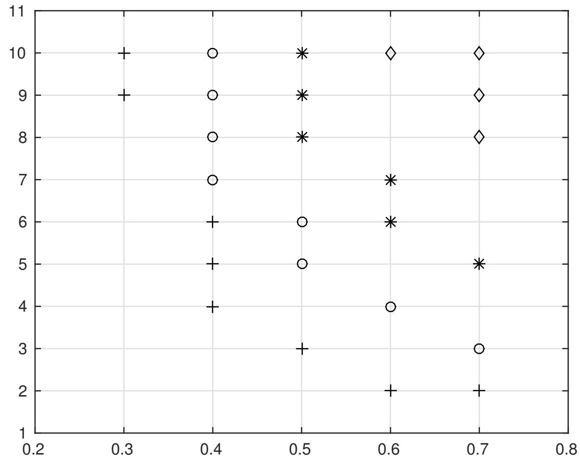

Figure 5: Plots of E_C as a function of $\lambda^2/\sqrt{n}$ (vertical axis) and $\varphi/\sqrt{n}$ (horizontal axis) for four ranges of E_C: $0.9 < E_C < 1.12 : +$; $1.13 < E_C < 1.16 : \circ$; $1.18 < E_C < 1.22 : *$; $E_C > 1.25 : \diamond$. Here $n = N^2 = 10^4$.

norm), which puts more weight on the diagonal elements. We derive exact formulas for the fitted SRE covariance parameters. Using a factorially designed study, we give regions of the factor space where the D-norm performs better than the F-norm. Specifically, it is better to use the D-norm, in terms of condition number, when the spatial dependence is strong.

Appendix

Proof of Proposition 1:

From Cressie and Johannesson (2008), let $\mathbf{C}$ be any positive-definite $n \times n$ matrix that plays the role of a target matrix. Recall that $\mathbf{S} = \mathbf{QR}$, and define $\mathbf{K}^* \equiv \mathbf{RKR}'$. Then $\mathbf{SKS}' = \mathbf{QK}^*\mathbf{Q}'$, and

$$\|\mathbf{C} - \mathbf{SKS}'\|_F^2 = \|\mathbf{C} - \mathbf{QK}^*\mathbf{Q}'\|_F^2 = \mathrm{tr}(\mathbf{C}'\mathbf{C}) + \mathrm{tr}\left((\mathbf{K}^*)'\mathbf{K}^*\right) - 2\mathrm{tr}(\mathbf{Q}'\mathbf{CQK}^*) \,. \tag{30}$$

Hence,

$$\frac{\partial}{\partial \mathbf{K}^*}\|\mathbf{C} - \mathbf{QK}^*\mathbf{Q}'\|_F^2 = 2\mathbf{K}^* - 2(\mathbf{Q}'\mathbf{CQ}) \,. \tag{31}$$

Putting this expression equal to the zero matrix yields $\mathbf{K}^* = \mathbf{Q}'\mathbf{CQ}$, which is positive-definite since $\mathbf{C}$ is positive-definite. Hence, $\hat{\mathbf{K}} \equiv \mathbf{R}^{-1}\mathbf{Q}'\mathbf{CQ}(\mathbf{R}^{-1})'$ is the estimate of $\mathbf{K}$ that minimizes $\|\mathbf{C} - \mathbf{SKS}'\|_F^2$. Now for a given σ_ξ^2, the previous result is applied to $\mathbf{C} = \mathbf{\Sigma}_0 - \sigma_\xi^2\mathbf{I}_n$. We define

$$\mathbf{K}(\sigma_\xi^2) \equiv \mathbf{R}^{-1}\mathbf{Q}'(\mathbf{\Sigma}_0 - \sigma_\xi^2\mathbf{I}_n)\mathbf{Q}(\mathbf{R}^{-1})' \,. \tag{32}$$

Then the minimum F-norm estimator of $\boldsymbol{\theta} = (\mathbf{K}, \sigma_\xi^2)$ is given by,

$$\hat{\sigma}_\xi^2 \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} \|\mathbf{\Sigma}_0 - \mathbf{SK}(\sigma_\xi^2)\mathbf{S}' - \sigma_\xi^2\mathbf{I}_n\|_F, \tag{33}$$

$$\hat{\mathbf{K}} \equiv \mathbf{K}(\hat{\sigma}_\xi^2). \tag{34}$$

In equation(33), restriction of $\boldsymbol{\theta} \in \Theta$ means that $\sigma_\xi^2 > 0$ and $\mathbf{C} = \boldsymbol{\Sigma}_0 - \sigma_\xi^2 \mathbf{I}_n$ is positive-definite. The minimization in (33) is only with respect to σ_ξ^2 and can be obtained straightforwardly. To see this, use (32) and $\mathbf{S} = \mathbf{QR}$ to write $\boldsymbol{\Sigma}_0 - \mathbf{SK}(\sigma_\xi^2)\mathbf{S}' - \sigma_\xi^2 \mathbf{I}_n \equiv \mathbf{G} + \sigma_\xi^2 \mathbf{H}$ with $\mathbf{G} = \boldsymbol{\Sigma}_0 - \mathbf{QQ}'\boldsymbol{\Sigma}_0\mathbf{QQ}'$ and $\mathbf{H} = \mathbf{QQ}' - \mathbf{I}_n$. Then $\|\mathbf{G} + \sigma_\xi^2 \mathbf{H}\|_F^2 = \sum_{i=1}^n \sum_{j=1}^n (g_{ij} + \sigma_\xi^2 h_{ij})^2$, and its derivative with respect to σ_ξ^2 is $2 \sum_{i=1}^n \sum_{j=1}^n (g_{ij} + \sigma_\xi^2 h_{ij}) h_{ij}$; putting this equal to zero and solving for σ_ξ^2, one obtains,

$$\hat{\sigma}_\xi^2 = -\frac{\sum_{i=1}^n \sum_{j=1}^n \left((\boldsymbol{\Sigma}_0 - \mathbf{QQ}'\boldsymbol{\Sigma}_0\mathbf{QQ}') \circ (\mathbf{QQ}' - \mathbf{I}_n)\right)_{ij}}{\|\mathbf{QQ}' - \mathbf{I}_n\|_F^2} , \tag{35}$$

where $\mathbf{A} \circ \mathbf{B}$ denotes the Hadamard product of two matrices $\mathbf{A}$ and $\mathbf{B}$, that is $(\mathbf{A} \circ \mathbf{B})_{ij} = (\mathbf{A})_{ij} \times (\mathbf{B})_{ij}$. Let us note here that we can't have $\mathbf{QQ}' - \mathbf{I}_n = \mathbf{0}$, because the rank of $\mathbf{Q}$ is less than or equal to r. The expression above in (35) is the same as (12), with the numerator expressed in terms of the vec operator.

Proof of Proposition 2:

Let us recall (14):

$$\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}(\boldsymbol{\theta})\|_D^2 = \|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}(\boldsymbol{\theta})\|_F^2 + \lambda^2 \|\mathrm{diag}(\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}(\boldsymbol{\theta}))\|_F^2 .$$

Since we have already evaluated (and differentiated) the first term of the right-hand side in the proof of Proposition 1, we turn our attention to evaluating and differentiating the second term. We use the notations given in the proof of Proposition 1.

Initially, assume that $\sigma_\xi^2 = 0$; then,

$$\begin{aligned} \|\mathrm{diag}(\boldsymbol{\Sigma}_0 - \mathbf{SKS}')\|_F^2 &= \mathrm{tr}\left((\mathrm{diag}\boldsymbol{\Sigma}_0)(\mathrm{diag}\boldsymbol{\Sigma}_0)\right) + \mathrm{tr}\left((\mathrm{diag}(\mathbf{SKS}'))(\mathrm{diag}(\mathbf{SKS}'))\right) \\ &\quad - 2\mathrm{tr}\left((\mathrm{diag}\boldsymbol{\Sigma}_0)(\mathrm{diag}(\mathbf{SKS}'))\right) . \end{aligned} \tag{36}$$

From the Q-R decomposition, $\mathbf{S} = \mathbf{QR}$, and recall that $\mathbf{SKS}' = \mathbf{QK}^*\mathbf{Q}'$, where $\mathbf{K}^* = \mathbf{RKR}'$. Hence the right-hand side of (36) becomes,

$$\mathrm{tr}\left((\mathrm{diag}(\boldsymbol{\Sigma}_0))^2\right) + \mathrm{tr}\left((\mathrm{diag}(\mathbf{QK}^*\mathbf{Q}'))^2\right) - 2\mathrm{tr}\left((\mathrm{diag}(\boldsymbol{\Sigma}_0))(\mathrm{diag}(\mathbf{QK}^*\mathbf{Q}'))\right) . \tag{37}$$

Our objective is to differentiate this expression with respect to $\mathbf{K}^*$. Recall the expression (31), which we now write in terms of the vec operator. That is,

$$\mathrm{vec}\left(\frac{\partial}{\partial k_{ab}^*} \|\mathbf{C} - \mathbf{C}^*(\mathbf{K}^*)\|_F^2\right) = 2\mathrm{vec}(\mathbf{K}^*) - 2\mathrm{vec}(\mathbf{Q}'\mathbf{CQ}) , \tag{38}$$

where k_{ab}^* is the (a, b) element of the $r \times r$ matrix $\mathbf{K}^*$.

Analogously, we differentiate (37) with respect to k_{ab}^*, for $a, b = 1, \ldots, r$. The differential of the first term in (37) is zero. If we write the $n \times r$ orthonormal matrix $\mathbf{Q}$ as (q_{ia}), the second term in (37) is:

$$\sum_{i=1}^n \left[\sum_{a=1}^r \sum_{b=1}^r q_{ia} k_{ab}^* q_{ib}\right]^2 ;$$

its differential with respect to k^*_{ab} is then,

$$2(q_{1a}q_{1b},\ldots,q_{na}q_{nb})\sum_{a'=1}^{r}\sum_{b'=1}^{r}\begin{pmatrix}q_{1a'}q_{1b'}\\ \vdots \\ q_{na'}q_{nb'}\end{pmatrix}k^*_{a'b'}$$

$$= 2\left((\mathbf{Q}_i\mathbf{Q}_i')_{ab} : i = 1,\ldots,n\right)\begin{pmatrix}\text{vec}(\mathbf{Q}_1\mathbf{Q}_1')'\text{vec}(\mathbf{K}^*)\\ \vdots \\ \text{vec}(\mathbf{Q}_n\mathbf{Q}_n')'\text{vec}(\mathbf{K}^*)\end{pmatrix},$$

where $\mathbf{Q}' \equiv (\mathbf{Q}_1 \ldots \mathbf{Q}_n)$.

The third term in (37) is:

$$-2\sum_{i=1}^{n}\sigma^0_{ii}\sum_{a=1}^{r}\sum_{b=1}^{r}q_{ia}k^*_{ab}q_{ib},$$

where the target covariance matrix is written as $\mathbf{\Sigma}_0 \equiv (\sigma^0_{ij})$, and hence $\text{diag}(\mathbf{\Sigma}_0)$ has $\sigma^0_{11},\ldots,\sigma^0_{nn}$ down its diagonal. Its differential with respect to k^*_{ab} is:

$$-2\left((\mathbf{Q}_i\mathbf{Q}_i')_{ab} : i = 1,\ldots,n\right)\begin{pmatrix}\sigma^0_{11}\\ \vdots \\ \sigma^0_{nn}\end{pmatrix}.$$

Now combine all three differentials, taken with respect to $\{k^*_{ab} : a, b = 1,\ldots,r\}$, to obtain:

$$\begin{aligned}\text{vec}\left(\frac{\partial}{\partial k^*_{ab}}\|\text{diag}(\mathbf{\Sigma}_0 - \mathbf{QK}^*\mathbf{Q}')\|_F^2\right) &= 2\left(\text{vec}(\mathbf{Q}_1\mathbf{Q}_1'),\ldots,\text{vec}(\mathbf{Q}_n\mathbf{Q}_n')\right)\begin{pmatrix}\text{vec}(\mathbf{Q}_1\mathbf{Q}_1')'\\ \vdots \\ \text{vec}(\mathbf{Q}_n\mathbf{Q}_n')'\end{pmatrix}\text{vec}(\mathbf{K}^*)\\ &\quad - 2\left(\text{vec}(\mathbf{Q}_1\mathbf{Q}_1'),\ldots,\text{vec}(\mathbf{Q}_n\mathbf{Q}_n')\right)\begin{pmatrix}\sigma^0_{11}\\ \vdots \\ \sigma^0_{nn}\end{pmatrix} \\ &\equiv 2\mathbf{g}(\mathbf{Q})\text{vec}(\mathbf{K}^*) - 2\mathbf{h}(\mathbf{Q},\text{diag}(\mathbf{\Sigma}_0)),\end{aligned} \tag{39}$$

where $\mathbf{g}(\mathbf{Q})$ defined just above is an $r^2 \times r^2$ matrix and $\mathbf{h}(\mathbf{Q},\text{diag}(\mathbf{\Sigma}_0))$ defined just above is an r^2-dimensional vector. Then

$$\text{vec}\left(\left(\frac{\partial}{\partial k^*_{ab}}\|\mathbf{\Sigma}_0 - \mathbf{QK}^*\mathbf{Q}'\|_D^2\right)\right) = 2\text{vec}(\mathbf{K}^*) - 2\text{vec}(\mathbf{Q}'\mathbf{\Sigma}_0\mathbf{Q}) + \lambda^2\left(2\mathbf{g}(\mathbf{Q})\text{vec}(\mathbf{K}^*) - 2\mathbf{h}(\mathbf{Q},\text{diag}(\mathbf{\Sigma}_0))\right).$$

Setting the right-hand side equal to the r^2-dimensional zero vector, yields the minimum D-norm fit,

$$\text{vec}(\hat{\mathbf{K}}^*) = \left(\mathbf{I}_{r^2} + \lambda^2\mathbf{g}(\mathbf{Q})\right)^{-1}\left\{\text{vec}(\mathbf{Q}'\mathbf{\Sigma}_0\mathbf{Q}) + \lambda^2\mathbf{h}(\mathbf{Q},\text{diag}(\mathbf{\Sigma}_0))\right\}. \tag{40}$$

We now use (40) to derive the required result when $\sigma^2_\xi > 0$. Finally then, the minimum D-norm fit is, for a given λ^2:

$$\hat{\sigma}^2_\xi(\lambda^2) \equiv \arg\min_{\sigma^2_\xi > 0}\|\mathbf{\Sigma}_0 - \mathbf{S}\hat{\mathbf{K}}(\sigma^2_\xi;\lambda^2)\mathbf{S}' - \sigma^2_\xi\mathbf{I}_n\|_D^2, \tag{41}$$

and

$$\hat{\mathbf{K}}(\lambda^2) \equiv \hat{\mathbf{K}}(\hat{\sigma}^2_\xi(\lambda^2);\lambda^2). \tag{42}$$

Acknowledgments

Hardouin's research was conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01). Cressie's research was supported by the Australian Research Council Discovery Project no. DP150104576, and by the US National Science Foundation under NSF grant SES-1132031 funded through the NSF-Census Research Network (NCRN) program.

References

Cressie, N., Johannesson, G., 2006. Spatial prediction for massive data sets, in: Mastering the Data Explosion in the Earth and Environmental Sciences: Proceedings of the Australian Academy of Science Elizabeth and Frederick White Conference, Canberra, Australia, Australian Academy of Science, pp. 1-11.

Cressie, N., Johannesson, G., 2008. Fixed Rank Kriging for very large spatial data sets. Journal of the Royal Statistical Society. Series B. 70, 209-226.

Lindgren, F., Rue, H., Lindstrom, J.,2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society. Series B. 73, 423-498.

Matheron, G., 1962. Traité de Géostatistique Apppliquée, Tome I. Mémoires du Bureau de Recherches Géologiques et Minières, No. 14, Editions Technip, Paris.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., Sain, S., 2015. A multiresolution Gaussian process model for the analysis of large spatial datasets. Journal of Computational and Graphical Statistics. 24, 579-599.

Sampson, P. D., Guttorp, P., 1992. Nonparametric estimation of nonstationary spatial covariance structure. Journal of the American Statistical Association. 87, 108-119.

Wikle, C. K., 2010. Low rank representations for spatial processes, in: Gelfand, A., Diggle, P., Fuentes, M., Guttorp, P. (Eds.), Handbook of Spatial Statistics. Chapman and Hall. CRC Press, Boca Raton, FL, pp. 107-118.