



HAL
open science

A diagonally weighted matrix norm between two covariance matrices

Noel Cressie, Cécile Hardouin

► **To cite this version:**

Noel Cressie, Cécile Hardouin. A diagonally weighted matrix norm between two covariance matrices. *Spatial Statistics*, 2019, 29, pp.316-328. 10.1016/j.spasta.2019.01.001 . hal-03120788

HAL Id: hal-03120788

<https://hal.science/hal-03120788>

Submitted on 21 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

A Diagonally Weighted Matrix Norm Between Two Covariance Matrices

Noel Cressie^a, Cécile Hardouin^b

^aNIASRA, University of Wollongong, Australia

^bMODAL'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

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

7 Spatial statistics has become important in many applications, particularly in the earth and
8 environmental sciences. Better sensors, for example on satellites, have led to a rapid increase in
9 the size n of spatial data sets. Kriging (Matheron, 1962) is an optimal method of spatial prediction
10 that filters out noise and fills in gaps in the data, but the kriging equations involve the inverse of
11 the $n \times n$ data covariance matrix Σ . In general, the computations to obtain the kriging predictor and
12 kriging variance are not scalable, usually of $O(n^3)$. Solutions to this problem include reduced-
13 dimension methods (see Wikle, 2010, for a review) and the use of sparse precision matrices
14 (Lindgren et al., 2011; Nychka et al., 2015). One of the reduced-dimension methods is based
15 on the spatial random effects (SRE) model, which is a spatial process given by a random linear
16 combination of r known spatial basis functions, where r is fixed and relatively small (Cressie and
17 Johannesson, 2006, 2008). The resulting spatial prediction, called Fixed Rank Kriging (FRK),
18 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

19 The SRE class of spatial covariance matrices is chosen to illustrate the methodology pre-
 20 sented in this article. One way to estimate the SRE-model parameters is via an EM algorithm,
 21 which requires parametric (usually Gaussian) assumptions. Alternatively, the SRE-model pa-
 22 rameters can be estimated via minimizing a Frobenius matrix norm (Cressie and Johannesson,
 23 2008) which, in this article, we generalize to a diagonally weighted Frobenius norm.

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

37 2. The Frobenius norm and its diagonally weighted version

38 2.1. The Frobenius norm (F-norm)

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

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

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

46 Now, suppose we wish to fit $\boldsymbol{\theta}$ by minimizing the norm of the difference, $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}(\boldsymbol{\theta})$, where $\boldsymbol{\Sigma}_0$
 47 is a target covariance matrix and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is a covariance matrix depending on unknown parameters
 48 $\boldsymbol{\theta}$. 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.
 49 In other settings, if $\mathbf{Z} = (Z_1, \dots, Z_n)'$ is an n -dimensional spatial process, then suppose we model
 50 $\text{cov}(\mathbf{Z}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$; if \mathbf{Z} is observed independently m times, resulting in data $\mathbf{Z}_1, \dots, \mathbf{Z}_m$, then we
 51 could choose for $\boldsymbol{\Sigma}_0$ the non-parametric estimator,

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

52 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
 53 replicates $\{\mathbf{Z}_t : t = 1 \dots m\}$ (over time) to obtain $\boldsymbol{\Sigma}_0$ given by (2).

54 Suppose that the target covariance matrix Σ_0 is obtained from the data, for example $\hat{\Sigma}_m$ in
 55 (2). A least-F-norm estimator of covariance parameters, θ , is defined as:

$$\hat{\theta} \equiv \arg \min_{\theta \in \Theta} \|\Sigma_0 - \Sigma(\theta)\|_F^2, \quad (3)$$

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

We shall now separate the variances from the covariances. Define

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

where $\text{diag}(\mathbf{B})$ is a diagonal matrix with $\{(\mathbf{B})_{ii} : i = 1, \dots, n\}$ down the diagonal, and $\theta_v \in \Theta_v \subset \Theta$
 are parameters of $(\text{var}(Z_1), \dots, \text{var}(Z_n))'$. Then, when the target covariance matrix Σ_0 is obtained
 from the data, a least-F-norm estimator, $\hat{\theta}_v$, can be obtained by minimizing with respect to θ_v ,

$$\|\text{diag}(\Sigma_0) - \mathbf{V}(\theta_v)\|_F^2 = \text{tr}(\text{diag}(\Sigma_0 - \Sigma(\theta))' \text{diag}(\Sigma_0 - \Sigma(\theta))).$$

61 That is,

$$\hat{\theta}_v = \arg \min_{\theta_v \in \Theta_v} \|\text{diag}(\Sigma_0) - \mathbf{V}(\theta_v)\|_F^2. \quad (4)$$

62 2.2. A diagonally weighted Frobenius norm (D-norm)

63 Motivated by (3) and (4), we introduce a diagonally weighted Frobenius norm (D-norm),
 64 $\|\mathbf{A}\|_D$, through

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

65 where λ^2 is fixed and, hence, the D-norm depends on it. Note that it is straightforward to show
 66 that $\|\cdot\|_D$ defined by (5) satisfies all the properties of a norm. Consequently, for $\lambda^2 > 0$, $\|\Sigma_0 -$
 67 $\Sigma(\theta)\|_D^2$ puts more emphasis on matching the variances than the covariances. Once again, suppose
 68 that the target covariance matrix Σ_0 is obtained from the data. Then define the least-D-norm
 69 estimator of θ as follows:

$$\hat{\theta}(\lambda^2) \equiv \arg \min_{\theta \in \Theta} \|\Sigma_0 - \Sigma(\theta)\|_D^2, \quad (6)$$

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

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

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

75 3.1. The SRE model

Suppose that $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ are spatial data on a finite set of locations, $D \equiv \{\mathbf{s}_i : i = 1, \dots, n\} \subset \mathbb{R}^d$, in a d -dimensional Euclidean space. We write $\mathbf{Z} = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$, 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 $i = 1, \dots, n$. We posit the following decomposition for $Z(\cdot)$: For $\mathbf{s} \in D$,

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

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

where $\mathbf{X}(\mathbf{s})'\boldsymbol{\beta}$ is the large-scale spatial variation due to p covariates, $\mathbf{X}(\cdot) \equiv (X_1(\cdot), \dots, X_p(\cdot))'$, 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):

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

76 where $\mathbf{S}(\cdot) \equiv (S_1(\cdot), \dots, S_r(\cdot))'$ is a vector of pre-specified, known spatial basis functions; $\boldsymbol{\eta} \equiv$
 77 $(\eta_1, \dots, \eta_r)'$ is a vector of random effects with mean zero and positive-definite covariance matrix
 78 \mathbf{K} , and $\xi(\cdot)$ represents the fine-scale variation in the process $Y(\cdot)$. It is assumed that $\xi(\cdot)$ has mean
 79 zero and correlation zero at distinct locations. That is, $\text{cov}(\xi(\mathbf{s}), \xi(\mathbf{u})) = \sigma_\xi^2 V(\mathbf{s})1(\mathbf{s} = \mathbf{u})$, where
 80 $\sigma_\xi^2 > 0$ is an unknown parameter, $V(\cdot) > 0$ is assumed known, and $1(\cdot)$ is an indicator function.
 81 Finally, $\xi(\cdot)$ is assumed to be statistically independent of $\boldsymbol{\eta}$.

82 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
 83 $Z(\cdot)$ is given by (7) and (8). Hence, we can assume that $\mathbf{X}(\cdot) \equiv \mathbf{0}$, since any fixed effect is ignored
 84 when calculating covariances. Then the model (8) reduces to

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

85 which in vector form can be written as

$$\mathbf{Z} = \mathbf{S}\boldsymbol{\eta} + \boldsymbol{\xi} + \boldsymbol{\varepsilon}, \quad (10)$$

86 where the three vectors on the right-hand side are mutually independent. In (10), $E(\boldsymbol{\eta}) = \mathbf{0}$
 87 and $\text{cov}(\boldsymbol{\eta}) = \mathbf{K}$; $E(\boldsymbol{\xi}) = \mathbf{0}$, and $\text{cov}(\boldsymbol{\xi}) = \sigma_\xi^2 \mathbf{V}$, where \mathbf{V} is a known diagonal matrix with
 88 $V(\mathbf{s}_1), \dots, V(\mathbf{s}_n)$ down the diagonal; and $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \mathbf{I}_n$, where \mathbf{I}_n is the n -
 89 dimensional identity matrix. Hence,

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbf{S}\mathbf{K}\mathbf{S}' + \sigma_\xi^2 \mathbf{V} + \sigma_\varepsilon^2 \mathbf{I}_n, \quad (11)$$

90 where $\boldsymbol{\theta} = (\mathbf{K}, \sigma_\xi^2)$. There is often an identifiability problem with estimating σ_ξ^2 and σ_ε^2 , which is
 91 resolved by assuming σ_ε^2 is known; we shall make that assumption here. In (11), parameters are
 92 $\boldsymbol{\theta} = (\mathbf{K}, \sigma_\xi^2) \in \Theta \equiv \{(\mathbf{K}, \sigma_\xi^2) : \mathbf{K} \text{ positive-definite, and } \sigma_\xi^2 > 0\}$.

93 3.2. Fitting SRE covariance parameters using the F-norm

94 The covariance parameters in the SRE model are given by \mathbf{K} and σ_ξ^2 in (11). For a target
 95 covariance matrix $\boldsymbol{\Sigma}_0$, we wish to fit $\boldsymbol{\theta} = (\mathbf{K}, \sigma_\xi^2)$ by minimizing the norm of the difference,
 96 $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}(\boldsymbol{\theta})$. Without loss of generality, we simplify (11) by putting $\sigma_\varepsilon^2 = 0$ and $\mathbf{V} = \mathbf{I}_n$. Otherwise,

our results still hold, albeit with more complicated formulas. Hence, our goal is to find $\hat{\theta} = (\hat{\mathbf{K}}, \hat{\sigma}_\xi^2) \in \Theta$, by minimizing $\|\Sigma_0 - \mathbf{S}\mathbf{K}\mathbf{S}' - \sigma_\xi^2 \mathbf{I}_n\|_F$; the restriction to the parameter space Θ means that $\hat{\mathbf{K}}$ is positive-definite and $\hat{\sigma}_\xi^2 > 0$. Write $\mathbf{S} = \mathbf{Q}\mathbf{R}$, 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 $\text{vec}(\mathbf{B}) \equiv (\mathbf{b}'_1 \mathbf{b}'_2 \dots \mathbf{b}'_n)'$ of the matrix $\mathbf{B} = (\mathbf{b}_1 \mathbf{b}_2 \dots \mathbf{b}_n)$.

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

Proposition 1. *Minimum F-norm estimator.*

Recall $(\hat{\mathbf{K}}, \hat{\sigma}_\xi^2) \equiv \arg \min_{\theta \in \Theta} \|\Sigma_0 - \mathbf{S}\mathbf{K}\mathbf{S}' - \sigma_\xi^2 \mathbf{I}_n\|_F^2$. Then

$$\hat{\sigma}_\xi^2 = \frac{(\text{vec}(\mathbf{Q}\mathbf{Q}'\Sigma_0\mathbf{Q}\mathbf{Q}' - \Sigma_0))' \text{vec}(\mathbf{Q}\mathbf{Q}' - \mathbf{I}_n)}{\|\mathbf{Q}\mathbf{Q}' - \mathbf{I}_n\|_F^2}, \quad (12)$$

and

$$\hat{\mathbf{K}} = \mathbf{R}^{-1} \mathbf{Q}' (\Sigma_0 - \hat{\sigma}_\xi^2 \mathbf{I}_n) \mathbf{Q} (\mathbf{R}^{-1})', \quad (13)$$

provided $\Sigma_0 - \hat{\sigma}_\xi^2 \mathbf{I}_n$ 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),

$$\|\Sigma_0 - \Sigma(\theta)\|_D^2 = \|\Sigma_0 - \Sigma(\theta)\|_F^2 + \lambda^2 \|\text{diag}(\Sigma_0 - \Sigma(\theta))\|_F^2, \quad (14)$$

where recall from (11) that $\Sigma(\theta) = \mathbf{S}\mathbf{K}\mathbf{S}' + \sigma_\xi^2 \mathbf{I}$, for $\theta = (\mathbf{K}, \sigma_\xi^2)$, \mathbf{K} positive-definite, and $\sigma_\xi^2 > 0$.

For λ^2 given, a least-D-norm estimate of θ is the parameter value that minimizes (14) above.

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

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

and

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

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:

$$\text{vec}(\hat{\mathbf{K}}^*(\sigma_\xi^2; \lambda^2)) \equiv (\mathbf{I}_{r^2} + \lambda^2 \mathbf{g}(\mathbf{Q}))^{-1} \left\{ \text{vec}(\mathbf{Q}'(\Sigma_0 - \sigma_\xi^2 \mathbf{I}_n)\mathbf{Q}) + \lambda^2 \mathbf{h}(\mathbf{Q}, \text{diag}(\Sigma_0 - \sigma_\xi^2 \mathbf{I}_n)) \right\}, \quad (17)$$

and hence define

$$\hat{\mathbf{K}}(\sigma_\xi^2; \lambda^2) \equiv \mathbf{R}^{-1} \hat{\mathbf{K}}^*(\sigma_\xi^2; \lambda^2) (\mathbf{R}^{-1})'. \quad (18)$$

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

120 **Proposition 2.** *Minimum D-norm estimator.*

121 For a given λ^2 , $\hat{\theta}(\lambda^2) \equiv \arg \min_{\theta \in \Theta} \|\Sigma_0 - \mathbf{S}\mathbf{K}\mathbf{S}' - \sigma_\xi^2 \mathbf{I}_n\|_D^2$ is given by

$$\hat{\sigma}_\xi^2(\lambda^2) = \arg \min_{\sigma_\xi^2 > 0} \|\Sigma_0 - \mathbf{S}\hat{\mathbf{K}}(\sigma_\xi^2; \lambda^2)\mathbf{S}' - \sigma_\xi^2 \mathbf{I}_n\|_D^2, \quad (19)$$

122 and

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

123 provided $\Sigma_0 - \hat{\sigma}_\xi^2(\lambda^2)\mathbf{I}_n$ is positive-definite.

124 Importantly, the minimization in (19) is restricted to those $\sigma_\xi^2 > 0$ that yield a positive-definite
 125 $\hat{\mathbf{K}}(\sigma_\xi^2; \lambda^2)$. From (17), this is guaranteed by considering only those $\sigma_\xi^2 > 0$ such that $\Sigma_0 - \sigma_\xi^2 \mathbf{I}_n$
 126 is positive-definite, which is the same condition given in Section 3.2 for the minimum F-norm
 127 estimator. Because of the closed-form expression for $\hat{\mathbf{K}}(\sigma_\xi^2; \lambda^2)$, the minimization in (19) is only
 128 with respect to the one-dimensional parameter $\sigma_\xi^2 > 0$, and it can be easily obtained by a golden
 129 search for example.

130 5. Application

131 In this section, we illustrate the advantage of using the D-norm in fitting an SRE model
 132 (9) to the well known exponential-covariance model, which is a particular case of the Matérn
 133 covariance model. We consider a two-dimensional lattice $D = \{\mathbf{s}_{ij} : i, j = 1, \dots, N\}$ with
 134 $N = 100$; that is, $n = 10^4$. We choose bisquare functions for the spatial basis functions, with
 135 three resolutions, the centers being regularly spaced within a resolution. The generic expression
 136 for these basis functions is,

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

137 where $\mathbf{c}_{j(l)}$ is the j th centre point of the l th resolution, for $l = 1, 2, 3$, and $\|\mathbf{s} - \mathbf{u}\|$ is the Euclidean
 138 distance between two locations \mathbf{s} and \mathbf{u} . The number of basis functions used at the three res-
 139 olutions are, respectively 5, 16, and 49. Consequently, the dimension of the reduced space is
 140 $r = 70$. The radius r_l of the l th resolution bisquare function equals 1.5 times the shortest distance
 141 between center points of this resolution, allowing overlap between the basis functions.

142 We want to find σ_ξ^2 and \mathbf{K} that minimize the norm of the difference, $\Sigma_0 - \Sigma(\sigma_\xi^2, \mathbf{K})$, where the
 143 target covariance $\Sigma_{0,ij} = C(\mathbf{s}_i, \mathbf{s}_j)$ is obtained from an exponential covariance function to which
 144 we choose to add a nugget effect. That is,

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

145 where c is the sill, φ is the scale parameter, and $a \geq 0$ is the nugget effect. Here we specify $c = 1$
 146 (without loss of generality), and φ ranges from 5 to 70, to capture weak to strong spatial depen-
 147 dence, respectively. We adopt this strategy because the spatial dependence in the exponential
 148 covariance function given by (21) is well understood. Our goal here is not parameter estimation,
 149 but it is to find σ_ξ^2 and \mathbf{K} that approximate the given covariance model Σ_0 with the “nearest” SRE
 150 covariance model.

151 We obtain $\hat{\mathbf{K}}_F$ and $\hat{\sigma}_{\xi,F}^2$ defined in (12) and (13), by minimizing $\|\Sigma_0 - \Sigma(\sigma_{\xi}^2, \mathbf{K})\|_F$; and we
 152 obtain $\hat{\mathbf{K}}_D(\lambda^2)$ and $\hat{\sigma}_{\xi,D}^2(\lambda^2)$ defined in (19) and (20), by minimizing $\|\Sigma_0 - \Sigma(\sigma_{\xi}^2, \mathbf{K})\|_D$, for various
 153 choices of λ^2 .

154 To compare the accuracy of the fits obtained from using the F-norm and the D-norm, we
 155 use a number of measures. Recall the Kullback-Leibler divergence, $D_{KL}(P_0|Q)$, where P_0 is a
 156 Gaussian distribution with mean $\mathbf{0}$ and covariance matrix Σ_0 , and Q is a Gaussian distribution of
 157 the same dimension with mean $\mathbf{0}$ and covariance matrix Σ_Q , as follows:

$$D_{KL}(P_0|Q) \equiv -\frac{1}{2} \log \left(\frac{\det \Sigma_0}{\det \Sigma_Q} \right) - \frac{n}{2} + \frac{1}{2} \text{tr}(\Sigma_Q^{-1} \Sigma_0) . \quad (22)$$

In our use of the Kullback-Leibler divergence in (22), Σ_Q is one or other of

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

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

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

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

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

$$E_2 \equiv \frac{\|\Sigma_0 - \hat{\Sigma}_F\|_2}{\|\Sigma_0 - \hat{\Sigma}_D(\lambda^2)\|_2} , \quad (26)$$

161 and

$$E_{\max} \equiv \frac{\|\Sigma_0 - \hat{\Sigma}_F\|_{\max}}{\|\Sigma_0 - \hat{\Sigma}_D(\lambda^2)\|_{\max}} , \quad (27)$$

162 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} .
 163 The following inequality holds between the norms we consider:

$$\|\cdot\|_{\max} \leq \|\cdot\|_2 \leq \|\cdot\|_F \leq \|\cdot\|_D . \quad (28)$$

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

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

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

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

172 $p \in \{0, 1/10, 1/3, 1/2, 2/3, 9/10\}$. The scale parameter $\varphi \in \Phi \equiv \{5, 10, 20, 30, 40, 50, 60, 70\}$;
 173 as φ increases from 5 to 70, it induces weak to strong spatial dependence. Finally, for the weights
 174 on the diagonal for the D-norm, we used smaller weights, $\lambda^2 \in \Lambda_1 \equiv \{0.1, 10, 20, 30, \dots, 100\}$, ,
 175 and larger weights, $\lambda^2 \in \Lambda_2 \equiv \{100k : k = 1, 2, \dots, 10\}$.

176 We now summarize the results obtained. First, the nugget effect does not impact the values
 177 of the ratios E_{max} , E_2 , E_C , and only very slightly those of E_{KL} . Hence, we choose to present the
 178 following results with $a = 0$, and we have chosen to compare results here for scale parameter
 179 $\varphi \in \{5, 20, 40, 70\}$. Plots of E_{KL} and E_C against λ^2 are presented in Figure 1 and Figure 2; and
 180 plots of E_2 and E_{max} against λ^2 are presented in Figures 3 and 4. Figures 1 and 3 show the case
 181 $\lambda^2 \in \Lambda_1$, while Figures 2 and 4 show the case $\lambda^2 \in \Lambda_2$.

182 When limiting the comparison to how well the original covariance matrix Σ_0 is fitted, it is
 183 clear that the D-norm performs in a similar manner to the F-norm, since E_{KL} and E_2 remain very
 184 close to 1. We have $0.9598 \leq E_{KL} \leq 1$. The smallest value of E_{KL} is obtained for $p = 90\%$, $\varphi =$
 185 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 .
 186 Similarly, we always have $0.9924 \leq E_2 \leq 1.0015$.

187 Now, we highlight the advantage of the D-norm with respect to the max norm, $\|\cdot\|_{max}$, and the
 188 condition number of the matrix $\hat{\mathbf{K}}$. The ratios of E_{max} increase with φ and with λ^2 . The values
 189 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
 190 $\lambda^2 \geq 700$. So, globally we can say that the D-norm performs better than the F-norm with respect
 191 to the matrix norm $\|\cdot\|_{max}$. Let us now consider the values of E_C , which is defined in terms of the
 192 SRE model's covariance-matrix parameter. As before, the ratios of E_C increase with φ and λ^2 ; E_C
 193 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
 194 is quite important. Also, the ratio E_C increases with φ ; for instance, for $\lambda^2 = 500$, E_C increases
 195 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
 196 $\varphi \in \Phi$. While the D-norm condition number does not improve for weak spatial dependence, it
 197 becomes more and more efficient to use the D-norm as the spatial dependence strengthens.

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

202 Finally, we present an empirical way of choosing λ^2 in Figure 5, where we plot λ^2 / \sqrt{n} against
 203 φ / \sqrt{n} for different fixed ranges of E_C . We choose the relative condition number E_C , because the
 204 inverse of the matrix \mathbf{K} is directly involved in the kriging equations, and hence, it is important
 205 that \mathbf{K} not be ill-conditioned. We considered four ranges of values of E_C in Figure 5, namely
 206 $0.9 < E_C < 1.12$, $1.13 < E_C < 1.16$, $1.18 < E_C < 1.22$, and $E_C > 1.25$, resulting in "gains" of
 207 about 10 percent, 15 percent, 20 percent, and more than 25 percent, respectively. For each fixed
 208 range, we recorded for each value of φ / \sqrt{n} the values of λ^2 / \sqrt{n} ensuring that E_C belongs to that
 209 range. Our main observation is that we need large values of λ^2 when the spatial dependence is
 210 moderate, and we need smaller values of λ^2 when the spatial dependence is strong. While no
 211 expression is derived linking E_C , λ^2 , φ , and n , it can be seen that $\lambda^2 / \sqrt{n} \geq (\varphi / \sqrt{n})^{-2}$ ensures
 212 that $E_C \geq 1.1$.

213 6. Discussion

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

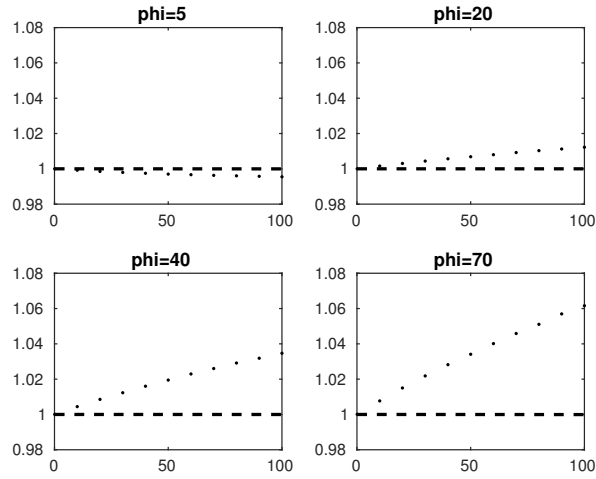


Figure 1: Plots of E_{KL} (—) and E_C (.) against $\lambda^2 \in \Lambda_1$ on the horizontal axis, for four values of $\varphi \in \Phi$.

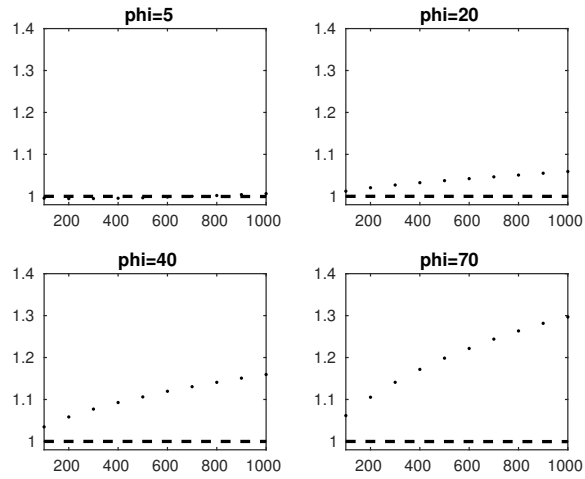


Figure 2: Plots of E_{KL} (—) and E_C (.) against $\lambda^2 \in \Lambda_2$ on the horizontal axis, for four values of $\varphi \in \Phi$.

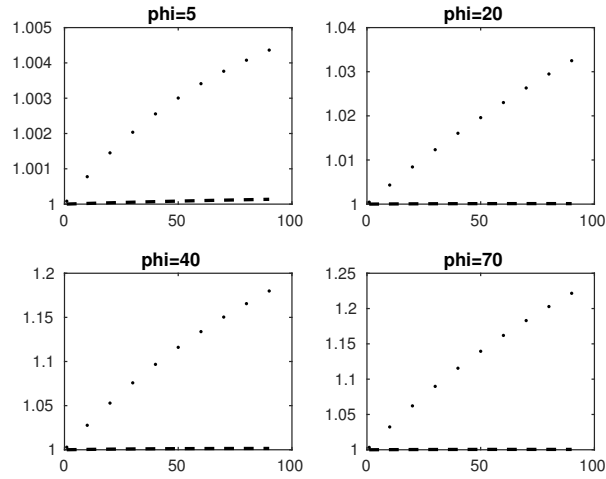


Figure 3: Plots of E_2 (–) and E_{max} (.) against $\lambda^2 \in \Lambda_1$ on the horizontal axis, for four values of $\varphi \in \Phi$.

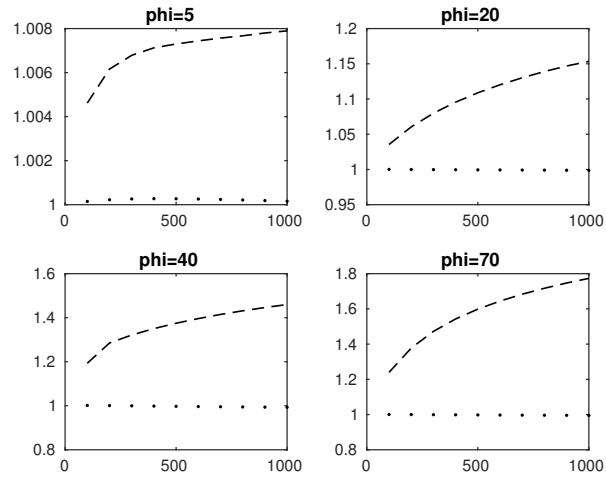


Figure 4: Plots of E_2 (–) and E_{max} (.) against $\lambda^2 \in \Lambda_2$ on the horizontal axis, for four values of $\varphi \in \Phi$.

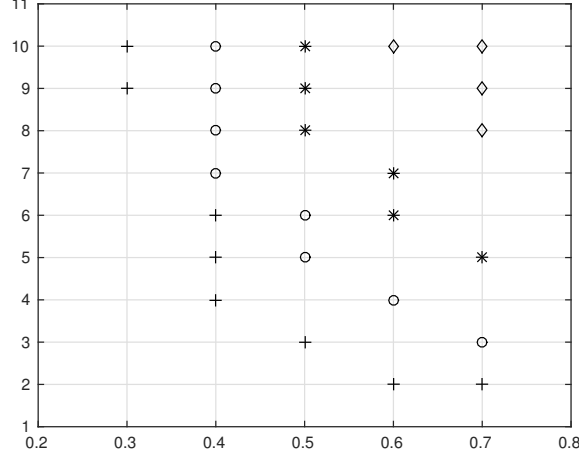


Figure 5: Plots of E_C as a function of λ^2/\sqrt{n} (vertical axis) and φ/\sqrt{n} (horizontal axis) for four ranges of E_C : $0.9 < E_C < 1.12$: +; $1.13 < E_C < 1.16$: o; $1.18 < E_C < 1.22$: *; $E_C > 1.25$: \diamond . Here $n = N^2 = 10^4$.

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

220 Appendix

221 *Proof of Proposition 1:*

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

$$\|\mathbf{C} - \mathbf{S}\mathbf{K}\mathbf{S}'\|_F^2 = \|\mathbf{C} - \mathbf{Q}\mathbf{K}^*\mathbf{Q}'\|_F^2 = \text{tr}(\mathbf{C}'\mathbf{C}) + \text{tr}((\mathbf{K}^*)'\mathbf{K}^*) - 2\text{tr}(\mathbf{Q}'\mathbf{C}\mathbf{Q}\mathbf{K}^*). \quad (30)$$

225 Hence,

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

226 Putting this expression equal to the zero matrix yields $\mathbf{K}^* = \mathbf{Q}'\mathbf{C}\mathbf{Q}$, which is positive-definite
 227 since \mathbf{C} is positive-definite. Hence, $\hat{\mathbf{K}} \equiv \mathbf{R}^{-1}\mathbf{Q}'\mathbf{C}\mathbf{Q}(\mathbf{R}^{-1})'$ is the estimate of \mathbf{K} that minimizes
 228 $\|\mathbf{C} - \mathbf{S}\mathbf{K}\mathbf{S}'\|_F^2$. Now for a given σ_ξ^2 , the previous result is applied to $\mathbf{C} = \boldsymbol{\Sigma}_0 - \sigma_\xi^2\mathbf{I}_n$. We define

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

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

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

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

229 In equation(33), restriction of $\theta \in \Theta$ means that $\sigma_\xi^2 > 0$ and $\mathbf{C} = \mathbf{\Sigma}_0 - \sigma_\xi^2 \mathbf{I}_n$ is positive-definite.
 230 The minimization in (33) is only with respect to σ_ξ^2 and can be obtained straightforwardly. To see
 231 this, use (32) and $\mathbf{S} = \mathbf{QR}$ to write $\mathbf{\Sigma}_0 - \mathbf{SK}(\sigma_\xi^2)\mathbf{S}' - \sigma_\xi^2 \mathbf{I}_n \equiv \mathbf{G} + \sigma_\xi^2 \mathbf{H}$ with $\mathbf{G} = \mathbf{\Sigma}_0 - \mathbf{QQ}'\mathbf{\Sigma}_0\mathbf{QQ}'$
 232 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
 233 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,

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

234 where $\mathbf{A} \circ \mathbf{B}$ denotes the Hadamard product of two matrices \mathbf{A} and \mathbf{B} , that is $(\mathbf{A} \circ \mathbf{B})_{ij} = (\mathbf{A})_{ij} \times$
 235 $(\mathbf{B})_{ij}$. 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
 236 equal to r . The expression above in (35) is the same as (12), with the numerator expressed in
 237 terms of the vec operator.

238 *Proof of Proposition 2:*

Let us recall (14):

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

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

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

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

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

$$\text{tr}((\text{diag}(\mathbf{\Sigma}_0))^2) + \text{tr}((\text{diag}(\mathbf{QK}^*\mathbf{Q}'))^2) - 2\text{tr}((\text{diag}(\mathbf{\Sigma}_0) (\text{diag}(\mathbf{QK}^*\mathbf{Q}')))). \quad (37)$$

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

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

247 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 $a, b = 1, \dots, r$. 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,

$$\begin{aligned} & 2(q_{1a}q_{1b}, \dots, q_{na}q_{nb}) \sum_{a'=1}^r \sum_{b'=1}^r \begin{pmatrix} q_{1a'}q_{1b'} \\ \vdots \\ q_{na'}q_{nb'} \end{pmatrix} k_{a'b'}^* \\ &= 2 \left((\mathbf{Q}_i \mathbf{Q}_i')_{ab} : i = 1, \dots, n \right) \begin{pmatrix} \text{vec}(\mathbf{Q}_1 \mathbf{Q}_1')' \text{vec}(\mathbf{K}^*) \\ \vdots \\ \text{vec}(\mathbf{Q}_n \mathbf{Q}_n')' \text{vec}(\mathbf{K}^*) \end{pmatrix}, \end{aligned}$$

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

The third term in (37) is:

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

where the target covariance matrix is written as $\Sigma_0 \equiv (\sigma_{ij}^0)$, and hence $\text{diag}(\Sigma_0)$ has $\sigma_{11}^0, \dots, \sigma_{nn}^0$ down its diagonal. Its differential with respect to k_{ab}^* is:

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

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

$$\begin{aligned} \text{vec} \left(\frac{\partial}{\partial k_{ab}^*} \|\text{diag}(\Sigma_0 - \mathbf{Q} \mathbf{K}^* \mathbf{Q}')\|_F^2 \right) &= 2 \left(\text{vec}(\mathbf{Q}_1 \mathbf{Q}_1'), \dots, \text{vec}(\mathbf{Q}_n \mathbf{Q}_n') \right) \begin{pmatrix} \text{vec}(\mathbf{Q}_1 \mathbf{Q}_1')' \\ \vdots \\ \text{vec}(\mathbf{Q}_n \mathbf{Q}_n')' \end{pmatrix} \text{vec}(\mathbf{K}^*) \\ &\quad - 2 \left(\text{vec}(\mathbf{Q}_1 \mathbf{Q}_1'), \dots, \text{vec}(\mathbf{Q}_n \mathbf{Q}_n') \right) \begin{pmatrix} \sigma_{11}^0 \\ \vdots \\ \sigma_{nn}^0 \end{pmatrix} \\ &\equiv 2\mathbf{g}(\mathbf{Q}) \text{vec}(\mathbf{K}^*) - 2\mathbf{h}(\mathbf{Q}, \text{diag}(\Sigma_0)), \end{aligned} \quad (39)$$

where $\mathbf{g}(\mathbf{Q})$ defined just above is an $r^2 \times r^2$ matrix and $\mathbf{h}(\mathbf{Q}, \text{diag}(\Sigma_0))$ defined just above is an r^2 -dimensional vector. Then

$$\text{vec} \left(\left(\frac{\partial}{\partial k_{ab}^*} \|\Sigma_0 - \mathbf{Q} \mathbf{K}^* \mathbf{Q}'\|_D^2 \right) \right) = 2\text{vec}(\mathbf{K}^*) - 2\text{vec}(\mathbf{Q}' \Sigma_0 \mathbf{Q}) + \lambda^2 (2\mathbf{g}(\mathbf{Q}) \text{vec}(\mathbf{K}^*) - 2\mathbf{h}(\mathbf{Q}, \text{diag}(\Sigma_0))).$$

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

$$\text{vec}(\hat{\mathbf{K}}^*) = (\mathbf{I}_{r^2} + \lambda^2 \mathbf{g}(\mathbf{Q}))^{-1} \left\{ \text{vec}(\mathbf{Q}' \Sigma_0 \mathbf{Q}) + \lambda^2 \mathbf{h}(\mathbf{Q}, \text{diag}(\Sigma_0)) \right\}. \quad (40)$$

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

$$\hat{\sigma}_\xi^2(\lambda^2) \equiv \arg \min_{\sigma_\xi^2 > 0} \|\Sigma_0 - \mathbf{S} \hat{\mathbf{K}}(\sigma_\xi^2; \lambda^2) \mathbf{S}' - \sigma_\xi^2 \mathbf{I}_n\|_D^2, \quad (41)$$

254 and

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

255 **Acknowledgments**

256 Hardouin's research was conducted as part of the project Labex MME-DII (ANR11-LBX-
257 0023-01). Cressie's research was supported by the Australian Research Council Discovery
258 Project no. DP150104576, and by the US National Science Foundation under NSF grant SES-
259 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 Appliquée*, 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.