

Asymptotic properties of multivariate tapering for estimation and prediction

Reinhard Furrer, François Bachoc, Juan Du

► **To cite this version:**

Reinhard Furrer, François Bachoc, Juan Du. Asymptotic properties of multivariate tapering for estimation and prediction. 2015. <hal-01203600>

HAL Id: hal-01203600

<https://hal.archives-ouvertes.fr/hal-01203600>

Submitted on 24 Sep 2015

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.

Asymptotic properties of multivariate tapering for estimation and prediction

Reinhard Furrer
reinhard.furrer@math.uzh.ch
University of Zurich

François Bachoc
francois.bachoc@univie.ac.at
University of Vienna

Juan Du
dujuan@ksu.edu
Kansas State University

June 8, 2015

Abstract: Parameter estimation for and prediction of spatially or spatio-temporally correlated random processes are used in many areas and often require the solution of a large linear system based on the covariance matrix of the observations. In recent years, the dataset sizes to which these methods are applied have steadily increased such that straightforward statistical tools are computationally too expensive to be used. In the univariate context, tapering, i.e., creating sparse approximate linear systems, has been shown to be an efficient tool in both the estimation and prediction settings. The asymptotic properties are derived under an infill asymptotic setting. In this paper we use a domain increasing framework for estimation and prediction using multivariate tapering. Under this asymptotic regime we prove that tapering (one-tapered form) preserves the consistency of the untapered maximum likelihood estimator and show that tapering has asymptotically the same mean squared prediction error as using the corresponding untapered predictor. The theoretical results are illustrated with simulations.

Keywords: one-taper likelihood; Gaussian random field; domain increasing; sparse matrix.

1 Introduction

Parameter estimation for and smoothing or interpolation of spatially or spatio-temporally correlated random processes are used in many areas and often require the solution of a large linear system based on the covariance matrix of the observations. In recent years, the dataset sizes to which these methods are applied have steadily increased such that straightforward statistical tools are computationally too expensive to be used. For example, a typical Landsat 7 satellite image consists of more than 34 million pixels (30 m resolution for an approximate scene size of 170 km \times 183 km; source landsat.usgs.gov). Hence, classical spatial and spatio-temporal models for such data sizes cannot be handled with typical soft- and hardware. Thus, one typically relies on approximation approaches. In the univariate context, tapering, i.e. creating sparse approximate linear systems through a direct product of the (presumed) covariance function and a positive definite but compactly supported correlation function, has been shown to be an efficient tool in both the estimation and prediction settings.

The vast majority of the theoretical work on univariate tapering has been placed in an infill-asymptotic setting using the concept of Gaussian equivalent measures and mis-specified covariance functions set forth in a series of papers by M. Stein (1988; 1990; 1997; 1999). Subsequently, Furrer *et al.* (2006); Kaufman *et al.* (2008); Du *et al.* (2009) and Wang and Loh (2011) have assumed a second-order stationary and isotropic Matérn covariance to show asymptotic optimality for prediction, consistency, and asymptotic efficiency for estimation. Recently, Stein

(2013) has extended these results to other covariance functions by placing appropriate conditions on the spectral density of the covariance.

In the infill–asymptotic setting, it is (essentially) sufficient to match the degree of differentiability at the origin of an appropriately chosen taper function with the smoothness of the (Matérn) covariance at the origin. Loosely speaking, for prediction, the predictor based on tapered covariances has the same convergence rate as the optimal predictor and the naive formula for the prediction kriging variance has the correct convergence rate as well (Theorem 2.1 of Furrer *et al.*, 2006, Theorem 1 of Stein, 2013).

For estimation, Kaufman *et al.* (2008) introduced the concept of one-taper and two-taper likelihood equations. In a one-taper setting only the covariance is tapered while for two-tapered both the covariance and empirical covariance are affected. The one-taper equation results in biased estimates while the two-taper equation is an estimating equation approach and is thus unbiased. The price of unbiased estimates is a (severe) loss of the computational efficiency intended through tapering (see, e.g., Table 2 of Kaufman *et al.*, 2008 or Figure 2 of Shaby and Ruppert, 2012).

Extending the idea of tapering to a multivariate setting is not straightforward. The infill–asymptotic setting does not allow one to ‘embed’ the multivariate framework in a univariate one (e.g., as in Sain *et al.*, 2011 for Gaussian Markov random fields). Ruiz-Medina and Porcu (2015) introduced the concept of multivariate Gaussian equivalent measures, but the conditions are difficult to verify and their practical applicability is not entirely convincing. Several authors have recently approached the problem using an increasing-domain setting (Shaby and Ruppert, 2012; Bevilacqua *et al.*, 2015). The main advantage of this alternative sampling scheme is that we are not bound to Matérn type covariance functions nor to tapers that satisfy the taper condition (i.e., sufficiently differentiable at the origin and at the taper length). More so, we will show that for collocated data, other practical tapers can be described. The main disadvantage is the somewhat less-intuitive conceptual framework. For example, in the case of heavy metal contents in sediments of a lake, infill–asymptotics can be mimicked by taking more and more measurements. In an increasing-domain setting, this is not possible. On the other hand asymptotics is a theoretical concept and in practice only a finite number of observations are available.

The main contributions of this paper are as follows: (i) under weak conditions on the covariance matrix function and the taper (matrix) function form we show that in an increasing-domain framework the tapered maximum likelihood estimator preserves the consistency of the untapered likelihood estimator; (ii) the difference between the (integrated) mean squared prediction error of the tapered and the untapered converges in probability to zero, even when prediction is based on estimated parameters. Note that although we require that the taper range increases, no rate assumption is necessary; (iii) numerical simulations illustrate that the approach has very appealing finite sample properties, especially for prediction with plugin estimates we find only a very small loss in efficiency.

This paper is structured as follows: Section 2 introduces basic notation and relevant definitions. The main results are given in Section 3. Section 4 illustrates the methodology using an extensive simulation study. Concluding remarks are given in Section 5. Proofs and technical results are presented in the appendix.

Note that compared with directly using compactly supported covariance functions, tapering has several advantages. Our modeling experience has shown that the (practical) dependence structure is often larger or much larger than what can be handled computationally and additional approximations would be needed anyway. We see tapering as a computational approximation

that does not alter the statistical model. The taper range (degree of tapering) depends on the availability of memory and computing power and thus changes when the analysis is carried out on different computers or at some later time with improved hardware.

2 Notation and setting

We denote (deterministic) vectors and matrices with bold lower and upper case symbols. Random variables and processes are denoted with upper case symbols and random vectors and vector processes are denoted with bold upper case symbols. For $\mathbf{x} \in \mathbb{R}^m$, we let $|\mathbf{x}| = \max_{i=1, \dots, m} |x_i|$ and $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^m x_i^2}$.

The singular values of a $n \times n$ real matrix $\mathbf{A} = (a_{ij})$ are denoted by $\rho_1(\mathbf{A}) \geq \dots \geq \rho_n(\mathbf{A}) \geq 0$ and, in the case when \mathbf{A} is symmetric, the eigenvalues are denoted by $\lambda_1(\mathbf{A}) \geq \dots \geq \lambda_n(\mathbf{A})$. The spectral norm is given by $\rho_1(\mathbf{A})$ and $\|\mathbf{A}\|_F^2 = \sum_{i,j} |a_{ij}|^2$ denotes the Frobenius norm.

For a sequence of random variables X_n , we write $X_n = o_p(1)$ when X_n converges to 0 in probability as $n \rightarrow \infty$ and we write $X_n = O_p(1)$ when X_n is bounded in probability as $n \rightarrow \infty$.

Let, for $d \in \mathbb{N}^+$ and $p \in \mathbb{N}^+$, fixed throughout this paper,

$$\{Z_k(\mathbf{s}) : \mathbf{s} \in \mathcal{D} \subset \mathbb{R}^d, k = 1, \dots, p\} \quad (1)$$

be a multivariate stationary Gaussian random process. We let $\mathbf{Z}(\mathbf{s}) = (Z_1(\mathbf{s}), \dots, Z_p(\mathbf{s}))^\top$. To simplify the notations, we assume, essentially without loss of generality, that:

Condition 1. *Process (1) has zero mean.*

Let $q \in \mathbb{N}^+$ and let Θ be the compact subset $[\theta_{\inf}, \theta_{\sup}]^q$ with $-\infty < \theta_{\inf} < \theta_{\sup} < +\infty$. For each $\boldsymbol{\theta} \in \Theta$ we consider a candidate stationary matrix covariance function for the process (1), of the form $\mathbf{C}(\mathbf{h}; \boldsymbol{\theta}) = (c_{kl}(\mathbf{h}; \boldsymbol{\theta}))$. We assume that there exists $\boldsymbol{\theta}_0 \in \Theta$, with for $i = 1, \dots, q$, $\theta_{\inf} < \theta_{0i} < \theta_{\sup}$, so that $\mathbf{C}(\mathbf{h}; \boldsymbol{\theta}_0) = \text{Cov}(\mathbf{Z}(\mathbf{s}), \mathbf{Z}(\mathbf{s} + \mathbf{h}))$. The covariance function $c_{kk}(\mathbf{h}; \boldsymbol{\theta}_0)$ of the k th (marginal) process is called a direct covariance (function) and the off-diagonal elements $c_{kl}(\mathbf{h}; \boldsymbol{\theta}_0)$, $k \neq l$, are called cross covariance (functions). We also consider a stationary taper matrix function of the form $(t_{kl}(\mathbf{h}))$, with $t_{kl}(\mathbf{h}) = 0$ for $\|\mathbf{h}\| \geq 1$.

For any $n \in \mathbb{N}^+$, the Gaussian processes (1) are observed at the points $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$:

Condition 2. *We dispose collocated observations at the distinct locations $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$.*

For $i = (k-1)n + a$ and $j = (l-1)n + b$, with $k, l = 1, \dots, p$ and $a, b = 1, \dots, n$, we let \mathbf{z} be the $np \times 1$ Gaussian vector with $z_i = Z_k(\mathbf{x}_a)$, for $\boldsymbol{\theta} \in \Theta$ we let $\boldsymbol{\Sigma}_\boldsymbol{\theta}$ be the $np \times np$ covariance matrix with $\sigma_{\boldsymbol{\theta}ij} = c_{kl}(\mathbf{x}_a - \mathbf{x}_b; \boldsymbol{\theta})$ and \mathbf{T} be the $np \times np$ taper covariance matrix with $t_{ij} = t_{kl}((\mathbf{x}_a - \mathbf{x}_b)/\gamma_n)$, where $\gamma_n > 0$ is the taper range. We let $\mathbf{K}_\boldsymbol{\theta} = \boldsymbol{\Sigma}_\boldsymbol{\theta} \circ \mathbf{T}$, where the symbol \circ denotes the direct (Schur) product.

The maximum likelihood (ML) estimator is defined by $\hat{\boldsymbol{\theta}}_{\text{ML}} \in \text{argmin}_\boldsymbol{\theta} L_\boldsymbol{\theta}$, with

$$L_\boldsymbol{\theta} = \frac{1}{np} \log(\det(\boldsymbol{\Sigma}_\boldsymbol{\theta})) + \frac{1}{np} \mathbf{z}^\top \boldsymbol{\Sigma}_\boldsymbol{\theta}^{-1} \mathbf{z}. \quad (2)$$

The tapered ML estimator is defined by $\hat{\boldsymbol{\theta}}_{t\text{ML}} \in \text{argmin}_\boldsymbol{\theta} \bar{L}_\boldsymbol{\theta}$, with

$$\bar{L}_\boldsymbol{\theta} = \frac{1}{np} \log(\det(\mathbf{K}_\boldsymbol{\theta})) + \frac{1}{np} \mathbf{z}^\top \mathbf{K}_\boldsymbol{\theta}^{-1} \mathbf{z}. \quad (3)$$

We can assume, without loss of generality, that $Z_1(\mathbf{x})$ is the Gaussian process that is predicted at new points. Then, for $\mathbf{x} \in \mathbb{R}^d$, let $\boldsymbol{\sigma}_\theta(\mathbf{x})$ be the $np \times 1$ vector defined by, for $i = (k-1)n + a$, $k = 1, \dots, p$, $a = 1, \dots, n$, $\sigma_\theta(\mathbf{x})_i = c_{1k}(\mathbf{x} - \mathbf{x}_a; \boldsymbol{\theta})$. Define similarly the $np \times 1$ vector $\mathbf{k}_\theta(\mathbf{x})$ by $k_\theta(\mathbf{x})_i = c_{1k}(\mathbf{x} - \mathbf{x}_a; \boldsymbol{\theta})t_{1k}((\mathbf{x} - \mathbf{x}_a)/\gamma_n)$.

3 Consistent estimation and asymptotically equal prediction

We first explore four conditions on covariance and taper matrix functions. The following condition holds for all the most classical models of covariance functions with infinite supports. Note that models with compactly supported covariance functions can be non-differentiable with respect to the covariance parameters, but that tapering is irrelevant anyway in increasing-domain asymptotics when the original covariance functions are already compactly supported.

Condition 3. For all fixed $\mathbf{x} \in \mathbb{R}^d$, $k, l = 1, \dots, p$, $c_{kl}(\mathbf{x}; \boldsymbol{\theta})$ is continuously differentiable with respect to $\boldsymbol{\theta}$. There exist constants $A < +\infty$ and $\alpha > 0$ so that for all $i = 1, \dots, q$, for all $\mathbf{x} \in \mathbb{R}^d$ and for all $\boldsymbol{\theta} \in \Theta$,

$$|c_{kl}(\mathbf{x}; \boldsymbol{\theta})| \leq \frac{A}{1 + |\mathbf{x}|^{d+\alpha}} \quad \text{and} \quad \left| \frac{\partial}{\partial \theta_i} c_{kl}(\mathbf{x}; \boldsymbol{\theta}) \right| \leq \frac{A}{1 + |\mathbf{x}|^{d+\alpha}}.$$

Condition 4. For all $k, l = 1, \dots, p$, the taper function t_{kl} is continuous at $\mathbf{0}$ and satisfies $t_{kl}(\mathbf{0}) = 1$ and $|t_{kl}(\mathbf{x})| \leq 1$ for all $\mathbf{x} \in \mathbb{R}^d$. The taper range $\gamma = \gamma_n$ satisfies $\gamma_n \rightarrow_{n \rightarrow \infty} +\infty$.

The next condition on a minimal distance between two different observation points is assumed in most domain increasing settings.

Condition 5. There exists a constant $\Delta > 0$ so that for all $n \in \mathbb{N}^+$ and for all $a \neq b$, $|\mathbf{x}_a - \mathbf{x}_b| \geq \Delta$.

Condition 6. There exists a constant $\delta > 0$ so that for all $n \in \mathbb{N}^+$ and for all $\boldsymbol{\theta} \in \Theta$, $\lambda_{np}(\boldsymbol{\Sigma}_\theta) \geq \delta$ and $\lambda_{np}(\mathbf{K}_\theta) \geq \delta$.

We expect Condition 6 to hold in many cases when Condition 5 also holds. For univariate tapering, Condition 6 would indeed hold under mild assumptions (consider an adaptation of Proposition D.4 in Bachoc, 2014b). Furthermore, when the parametric model incorporates a nugget effect or measurement errors, then Condition 6 holds provided that the nugget or error variances are lower-bounded uniformly in $\boldsymbol{\theta}$. The nugget or measurement error case is directly treated by Theorem 1; Theorem 3 would also be valid for it with a minor change of notation to define the integrated prediction errors (see, e.g., the context of Bachoc, 2014a).

The next theorem and corollary (the corollary is proved using standard M -estimator techniques), show that if the standard conditions for consistency of the (untapered) ML estimator hold, then the tapering preserves this consistency, as long as $\gamma \rightarrow_{n \rightarrow \infty} +\infty$.

Theorem 1. Assume that Conditions 3, 4, 5, and 6 hold. Then, as $n \rightarrow \infty$,

$$\sup_{\boldsymbol{\theta} \in \Theta} |L_\theta - \bar{L}_\theta| = o_p(1).$$

Corollary 2. Consider the same setting as in Theorem 1. Assume that for all $\kappa > 0$ there exists $\epsilon > 0$ so that

$$\inf_{|\boldsymbol{\theta} - \boldsymbol{\theta}_0| \geq \kappa} L_\theta - L_{\boldsymbol{\theta}_0} \geq \epsilon + o_p(1),$$

where the $o_p(1)$ may depend on ϵ and κ and goes to 0 in probability as $n \rightarrow \infty$. Then, as $n \rightarrow \infty$,

$$\widehat{\boldsymbol{\theta}}_{ML} \rightarrow_p \boldsymbol{\theta}_0 \quad \text{and} \quad \widehat{\boldsymbol{\theta}}_{tML} \rightarrow_p \boldsymbol{\theta}_0.$$

Theorem 1 and Corollary 2 highlight the important difference between one-taper and two-taper ML in terms of asymptotics. One-taper approximation with fixed range γ and independent of n boils down to an incorrectly specified covariance model. Thus, with fixed γ , the tapered ML estimator would generally be inconsistent and would converge to the asymptotic minimizer of a Kullback–Leibler divergence (for the univariate case, see the discussion in Kaufman *et al.*, 2008, and also Watkins and Al-Boutiahi, 1990, or Bachoc, 2014a). Hence, assuming $\gamma \rightarrow \infty$ is necessary to prove consistency, which we do here. Note that, nevertheless, no rate needs to be specified. These facts also entail an exposition benefit for our paper: we simply have to show that the one-taper approximation does not damage the untapered ML estimator. The question of the consistency of this latter estimator can be treated in separate references, like Mardia and Marshall (1984) or Bachoc (2014b) for the univariate case. Especially, identifiability assumptions for the covariance model need not be discussed in our paper.

On the other hand, for the two-taper ML, consistency can be proved for a fixed γ , provided notably that the model of tapered covariance and cross-covariance functions is identifiable. (In particular, two different covariance parameters yield two different sets of tapered covariance and cross-covariance functions.) We refer to Shaby and Ruppert (2012) for a corresponding proof in the univariate case. (Actually, we believe that a global identifiability condition might be missing in Shaby and Ruppert (2012), stronger than assumption (B) in this reference, for it is not clear how to go from (S.29) to (S.30) in its supplementary material.) Hence, the difference between the asymptotic analysis of the untapered and two-taper ML estimators is more pronounced, since the latter estimator is a quasi-likelihood estimator in a covariance model different from the original one. This is why, in Shaby and Ruppert (2012), many assumptions, notably on identifiability, are restated independently of the untapered ML estimator.

These asymptotic considerations also correspond to practical aspects of the comparison between one- and two-taper equations. The latter can be employed with a smaller range γ than the former, which is beneficial, but on the other hand, requires the full inverse of a sparse matrix.

The following theorem shows that tapering has no asymptotic effect on prediction, uniformly in the covariance parameter $\boldsymbol{\theta}$. (Note that for prediction, there is no distinction between one and two-taper approximation.)

Theorem 3. *Assume that Conditions 3, 4, 5, and 6 hold. Let $(\mathbf{x}_{new,n})_{n \in \mathbb{N}^+}$ be a fixed sequence in \mathbb{R}^d . Then, as $n \rightarrow \infty$,*

$$\sup_{\boldsymbol{\theta} \in \Theta} \left| \left[\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x}_{new,n})^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} - Z_1(\mathbf{x}_{new,n}) \right]^2 - \left[\mathbf{k}_{\boldsymbol{\theta}}(\mathbf{x}_{new,n})^\top \mathbf{K}_{\boldsymbol{\theta}}^{-1} \mathbf{z} - Z_1(\mathbf{x}_{new,n}) \right]^2 \right| = o_p(1). \quad (4)$$

Assume furthermore that for any fixed $\boldsymbol{\theta}$, k and l , the functions $c_{kl}(\mathbf{x}; \boldsymbol{\theta})$ and $t_{kl}(\mathbf{x})$ are continuous. Let \mathcal{D}_n be a sequence of measurable subsets of \mathbb{R}^d with positive Lebesgue measures and let $f_n(\mathbf{x})$ be a sequence of continuous probability density functions on \mathcal{D}_n . Then, as $n \rightarrow \infty$,

$$\sup_{\boldsymbol{\theta} \in \Theta} \left| \int_{\mathcal{D}_n} \left[\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x})^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} - Z_1(\mathbf{x}) \right]^2 f_n(\mathbf{x}) d\mathbf{x} - \int_{\mathcal{D}_n} \left[\mathbf{k}_{\boldsymbol{\theta}}(\mathbf{x})^\top \mathbf{K}_{\boldsymbol{\theta}}^{-1} \mathbf{z} - Z_1(\mathbf{x}) \right]^2 f_n(\mathbf{x}) d\mathbf{x} \right| = o_p(1). \quad (5)$$

In (5), we assume continuity of the cross covariance, covariance and taper functions, and of $f_n(\mathbf{x})$ in order to define integrals in the L^2 sense. When $f_n(\mathbf{x})$ is constant on \mathcal{D}_n , Theorem 3

shows that tapering does not damage the mean integrated square prediction error over any sequence of prediction domains \mathcal{D}_n . Furthermore, in (4) and (5), the terms in the differences are typically bounded away from zero in probability, because of Condition 5 (consider for example Equation (10) in Proposition 5.2 of Bachoc, 2014b). (This would not hold only in degenerate cases when $\mathbf{x}_{\text{new},n}$ becomes arbitrarily close to an observation point or where $f_n(\mathbf{x})$ concentrates around an observation point.) Hence, also the ratio of (integrated) mean square prediction errors, between tapered and untapered predictions, converges to unity in general. Finally, because of the supremum over $\boldsymbol{\theta}$ in (4) and (5), Theorem 3 implies that the difference of tapered and untapered prediction errors goes to zero also when the predictions are obtained from any common estimator $\widehat{\boldsymbol{\theta}}$.

Remark: The condition $t_{kl}(\mathbf{0}) = 1$ in Condition 4 is necessary for Theorem 1. Indeed, it is typically needed in order to guarantee that $1/(np)\|\boldsymbol{\Sigma}_{\boldsymbol{\theta}} - \mathbf{K}_{\boldsymbol{\theta}}\|_F^2$ goes to zero. The latter is necessary for Theorem 1, as can be shown from the arguments in the proof of Proposition 3.1 in Bachoc (2014b). The condition $t_{kl}(\mathbf{0}) = 1$ should also be needed for Theorem 3, as is suggested by the second offline equation in Proposition 5.1 in Bachoc (2014b).

4 Simulations and illustrations

We now evaluate the finite sample performance of multivariate tapering with simulations. We consider a bivariate Gaussian isotropic process with Matérn type direct and cross-covariances

$$c_{kl}(\mathbf{x}; \boldsymbol{\theta}) = \frac{\sigma_{kl}^2}{2^{\nu_{kl}-1}\Gamma(\nu_{kl})} (\|\mathbf{x}\|/\rho_{kl})^{\nu_{kl}} \mathcal{K}_{\nu_{kl}}(\|\mathbf{x}\|/\rho_{kl}), \quad k, l = 1, 2 \quad (6)$$

where Γ is the Gamma function and \mathcal{K}_{ν} is the modified Bessel function of the second kind of order ν (Abramowitz and Stegun, 1970). To ensure positive definiteness, constraints on $\{\sigma_{kl}, \rho_{kl}, \nu_{kl}, k, l = 1, 2\}$ have to be imposed, see Gneiting *et al.* (2010). We use two different covariance models:

- (A) ranges: $\rho_{11} = 5, \rho_{12} = 3, \rho_{22} = 4$
sills: $\sigma_{11} = 1, \sigma_{12} = .6, \sigma_{22} = 1$
smoothness: $\nu_{11} = \nu_{12} = \nu_{22} = 1/2$
- (B) ranges: $\rho_{11} = 3, \rho_{12} = 3, \rho_{22} = 4$
sills: $\sigma_{11} = 1, \sigma_{12} = .7, \sigma_{22} = 1$
smoothness: $\nu_{11} = 3/2, \nu_{12} = 1, \nu_{22} = 1/2$

The smoothness parameters will not be estimated and are fixed. Hence, $\boldsymbol{\theta} = (\rho_{11}, \rho_{12}, \rho_{22}, \sigma_{11}, \sigma_{12}, \sigma_{22})^T$ and $q = 6$. The Matérn covariance functions satisfy Condition 3.

We consider the following taper matrix functions:

- (i) $t_{kl}(\mathbf{x}) = (1 - \|\mathbf{x}\|_+)^4 (1 + 4\|\mathbf{x}\|)$, $k, l = 1, 2$.
- (ii) $t_{kl}(\mathbf{x}) = (1 - \|\mathbf{x}\|_+)^6 (1 + 6\|\mathbf{x}\| + 35\|\mathbf{x}\|^2/3)$, $k, l = 1, 2$.
- (iii) $t_{kl}(\mathbf{x}) = (1 - \|\mathbf{x}\|_+)^2 (1 + \|\mathbf{x}\|/2)$, $k, l = 1, 2$.
- (iv) $t_{11}(\mathbf{x}) = (1 - \|\mathbf{x}\|_+)^5 (1 + 5\|\mathbf{x}\| + \|\mathbf{x}\|^2)$, $t_{12}(\mathbf{x}) = t_{21}(\mathbf{x}) = \sqrt{6/7} (1 - \|\mathbf{x}\|_+)^5 (1 + 5\|\mathbf{x}\| + \|\mathbf{x}\|^2)$,
 $t_{22}(\mathbf{x}) = (1 - \|\mathbf{x}\|_+)^5 (1 + 5\|\mathbf{x}\|)$.

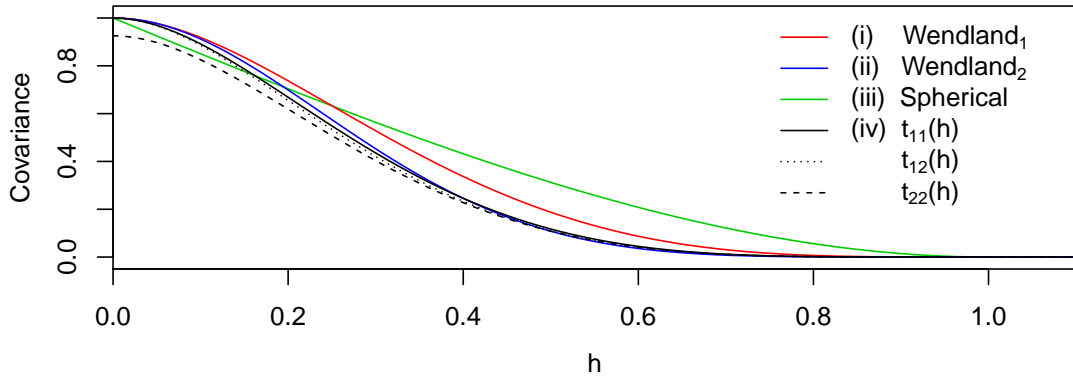


Figure 1: Different taper functions.

Taper matrix functions (i)–(iii) satisfy Condition 4 and the associated taper matrices are of the form $\mathbf{T} = \mathbf{1}\mathbf{1}^\top \otimes t(\|\mathbf{x}_a - \mathbf{x}_b\|/\gamma)$ where the symbol \otimes denotes the Kronecker product and where $t(\cdot)$ is as indicated above. In the literature these functions are referred to as Wendland₁, Wendland₂ and spherical taper (Wendland, 1995; Furrer *et al.*, 2006).

Taper matrix function (iv) is taken from Demel (2013) Corollary 2.2.3, based upon results from Theorem 3 of Ma (2011a) and Lemma 2 of Ma (2011b). The validity of this taper matrix function can also be shown using Theorem A in Daley *et al.* (2014) published later. Taper matrix function (iv) has $t_{12}(\mathbf{0}) = \sqrt{6/7} < 1$ (see Figure 1) and we investigate its finite sample behavior although Condition 4 is violated. We expect similar behavior of (i), (ii), and (iv) as the (direct) taper functions are very similar.

We are sampling $4m^2$ locations uniformly in a domain defined by the union of squares $[(1 - \Delta)/2]^2$, centered at $\{\pm(r - 1/2), \pm(s - 1/2)\}$, $r, s = 1, m$. The parameter Δ represents the minimum distance between the locations and the case $\Delta = 1$ is a regular grid. Prediction is done at the location $\mathbf{x}_{\text{new}} = (0, 0)^\top$ in the center of the domain. Figure 2 illustrates the setup. We present results for the two cases $\Delta = 0.2, 1$ (thus satisfying Condition 5) and three grid size parameter values $m = 10, 16, 25$, i.e., $n = 400, 1024, 2500$ and covariance matrix sizes 800×800 , 2048×2048 , 5000×5000 , respectively. Condition 6 has been verified numerically.

The next two subsections discuss the results of estimation and prediction. Computational details are given in the last subsection.

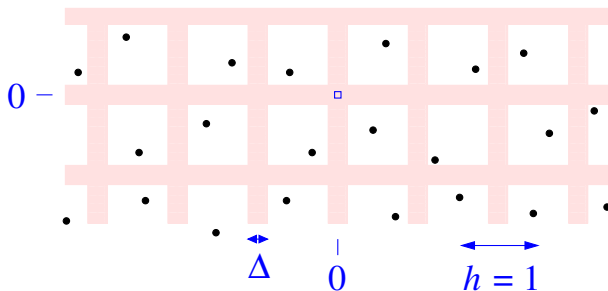


Figure 2: One set of sampled locations with simulation parameter $\Delta = 0.2$ and square center spacing $h = 1$.

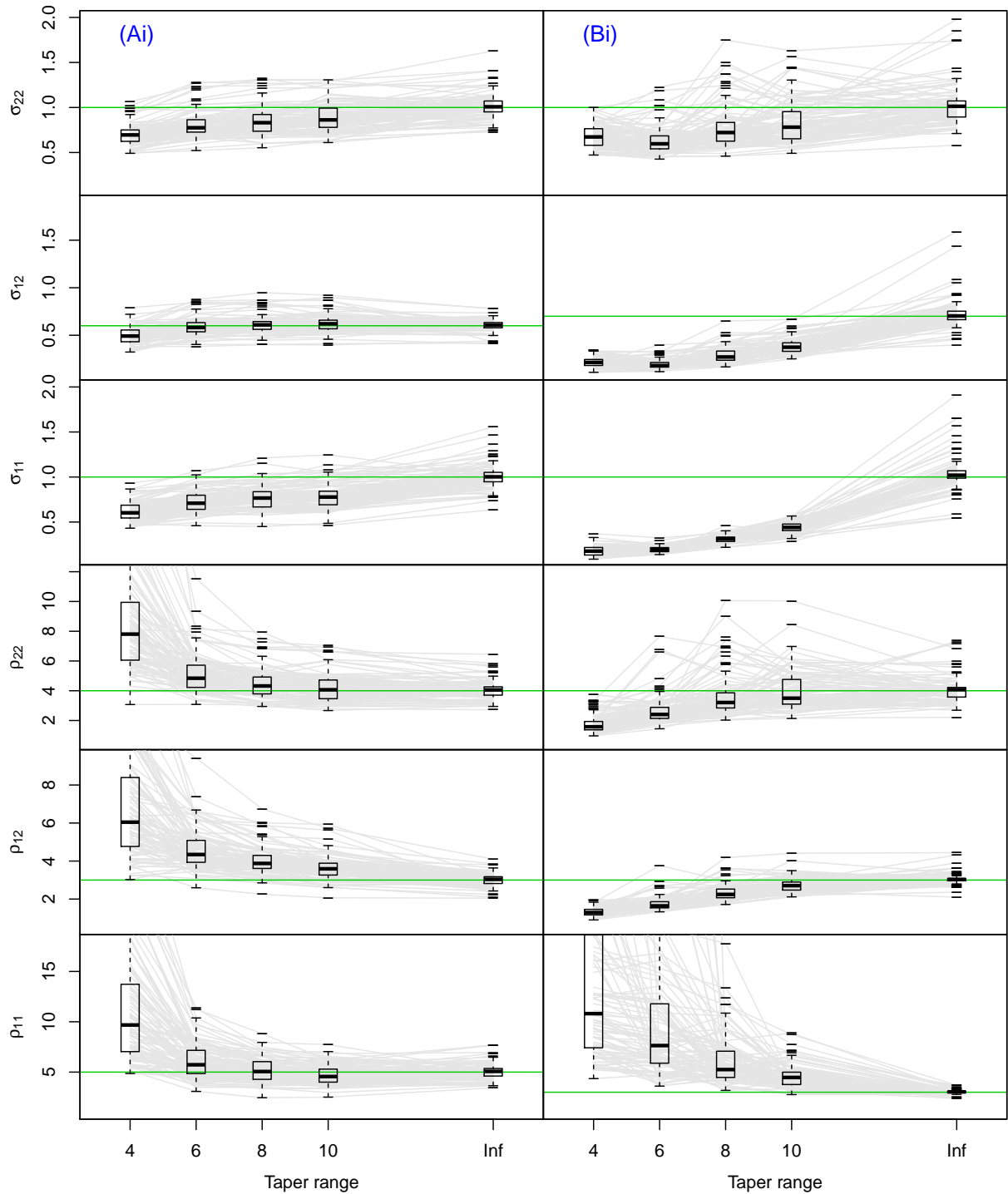


Figure 3: Effect of increasing the taper range γ on the ML estimates. Columns are for the two different covariance models, rows are for different parameters (truth is indicated by the horizontal green line). 100 realizations have been generated ($\Delta = 1$) based on $n = 400$. Each individual realization is indicated with a gray line.

4.1 Estimation

We first investigate $\hat{\theta}_{tML}$ and compare it to θ_0 as the taper range increases. Figure 3 summarizes the estimates of $\hat{\theta}_{tML}$ for equispaced observations ($\Delta = 1$) with $n = 400$, taper function (i), and using taper ranges $\gamma = 4, 6, 8, 10$ as well as no tapering ($\gamma = \text{Inf}$). As expected, for small taper ranges the results are biased with range parameters typically overestimated and sill parameters

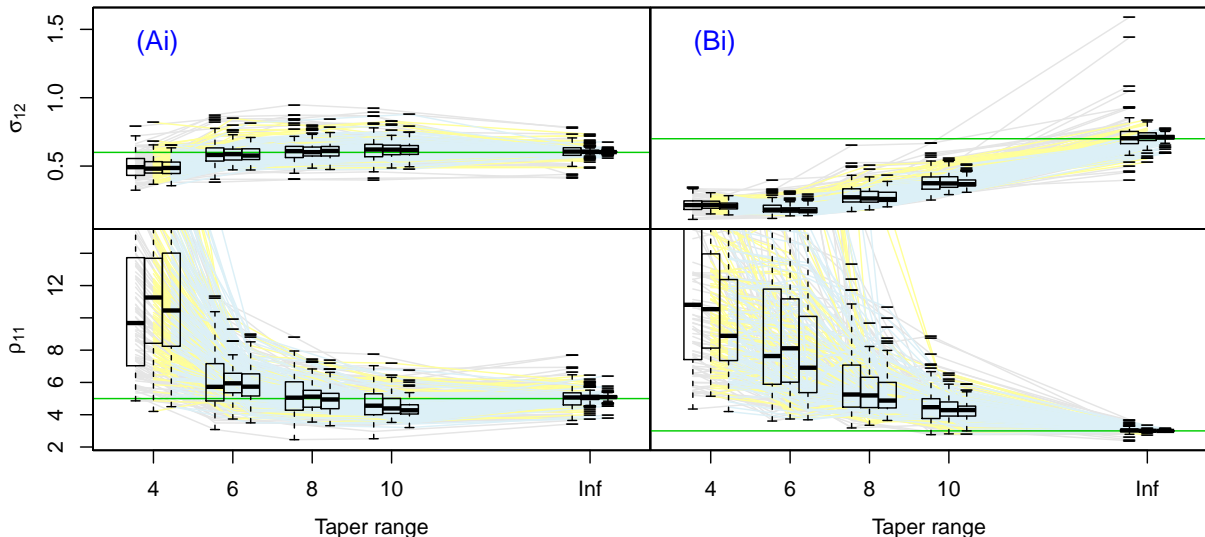


Figure 4: Effect of increasing the domain on the ML estimates. The boxplots correspond to $n = 400$ (gray), 1024 (yellow), 2500 (light blue), left to right for each taper range, $\Delta = 1$. See also Figure 3.

underestimated. For smoother spatial fields (B), the bias and uncertainties are (slightly) larger. The estimates of the sill parameters benefit from a regularizing aspect of tapering and thus exhibit a consistently smaller variance compared with the untapered estimates. This effect of regularizing is surprisingly strong for model (B) and parameter σ_{11} .

Figure 4 shows the effect of increasing the number of locations where we have added the boxplots for $n = 1024$ and $n = 2500$ (i.e., $m = 16$ and $m = 25$) to four panels of Figure 3. For the untapered estimates, one clearly sees that the uncertainties in the estimates decrease with increasing n . For the tapered estimates this effect is not as pronounced because of the regularizing effect of the tapering. As expected, the bias itself is not reduced by increasing the number of observations while keeping the taper range fixed. On the other hand, as illustrated in Corollary 2, when going from $n = 400, \gamma = 4$ to $n = 2500, \gamma = 10$, the distribution of the tapered ML estimates becomes closer to that of the untapered ones.

4.2 Prediction

In practice, prediction is often of prime interest and we primarily investigate the effect of tapering on the prediction of the first process Z_1 at the unobserved location $\mathbf{x}_{\text{new}} = (0, 0)^\top$. As parameter values we use $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_{t\text{ML}}$ for different taper ranges γ .

In Figure 5 we display the ratio of the tapered to the untapered mean squared prediction errors (MSPEs) using $\boldsymbol{\theta}_0$. For Model (A), the loss of efficiency is in general of the order of a few percent (the 95% pointwise range is below 1.08 for $\gamma \geq 5$). For smoother processes, the taper range needs to be increased in order to maintain the same efficiency. This is in sync with infill-asymptotic results (see, e.g., Figure 3 of Furrer *et al.*, 2006). There is little difference between the Wendland₁ and Wendland₂ tapers. Overall, the former having in general a slightly smaller MSPE.

The third row of Figure 5 illustrates why it is prohibitive to use tapers that are linear at the origin. While the spherical taper has no influence on the screening effect (Stein, 2002) of the exponential Model (A) (left panel) it completely breaks down for smoother fields (right panel).

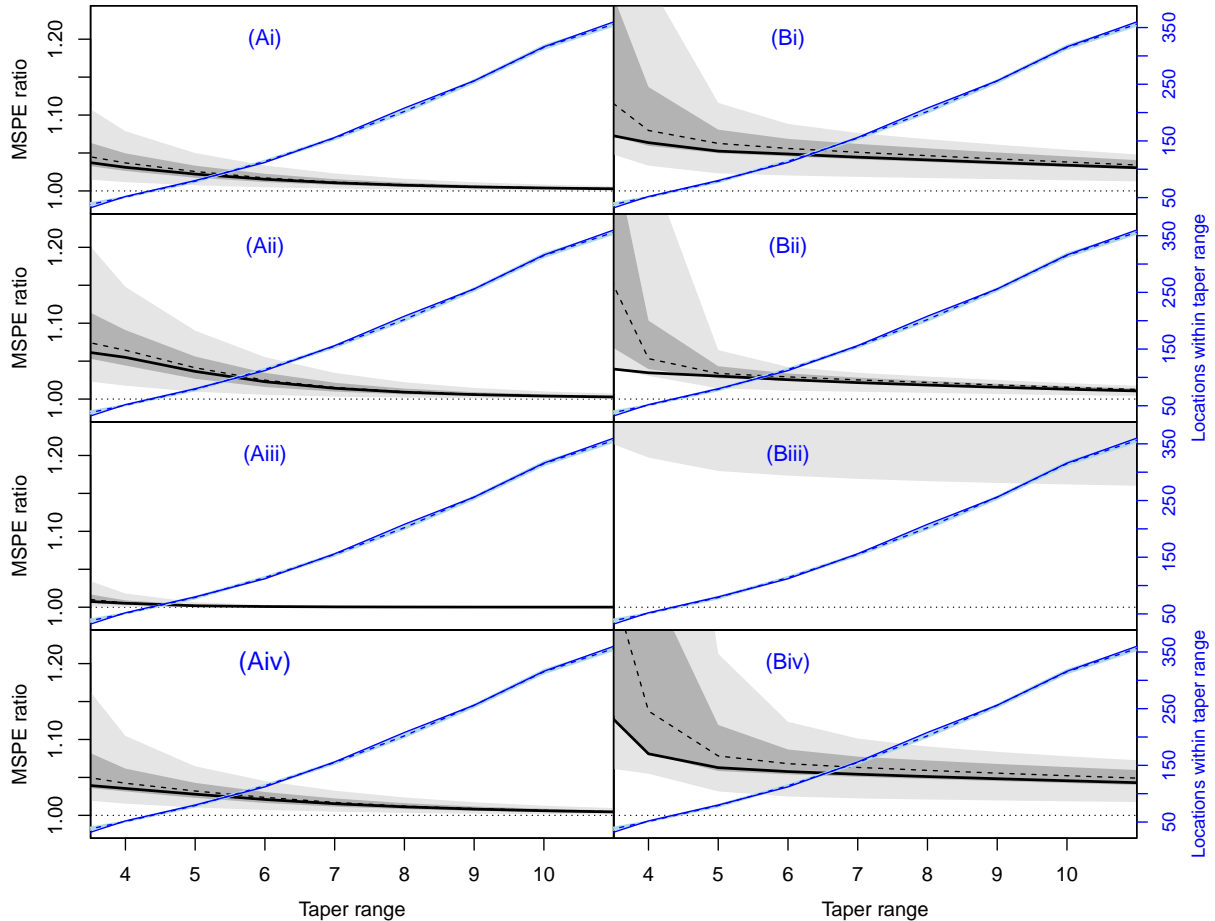


Figure 5: Ratios of the tapered to the untapered MSPEs for $n = 400$ using θ_0 . The solid line represents MSPE ratios for equispaced locations ($\Delta = 1$), the dashed line shows the median MSPE ratios from 100 simulations with random locations with $\Delta = 0.2$ (gray and light gray are pointwise 50 and 95 percentiles). The blue lines indicate the number of points within the taper range (mean solid, median dashed and light blue pointwise 95 percentiles).

Figure 5 also links the taper range with the number of observations within the taper range. The MSPE ratios suggest that tapering with more than 100 locations within the taper range is hardly worth the effort.

In Figure 5, we distinguish a small loss of efficiency when using taper function (iv) compared with (i) and (ii). This can be explained by the fact that the taper function (iv) does not satisfy Condition 4 (as $t_{12}(\mathbf{0}) < 1$). Nevertheless, this loss is far less pronounced than when using taper function (iii) for model (B).

For very small taper ranges, the MSPE ratios shown in Figure 5 seem large. However, presented in terms of differences, the effect of tapering is hardly noticeable. For example, for the setting (Ai) with $n = 400$, the MSPEs are 0.1155 0.1101 0.1098 for $\gamma = 3, 11, \infty$, respectively (see also red line in the left panel of Figure 6).

The left panel of Figure 6 further shows the effect of increasing the number of locations on the MSPE. The effect of increasing n is negligible even for the theoretical MSPE, the values are visually indistinguishable. With as few as $n = 400$ we extract essentially all the information in the system.

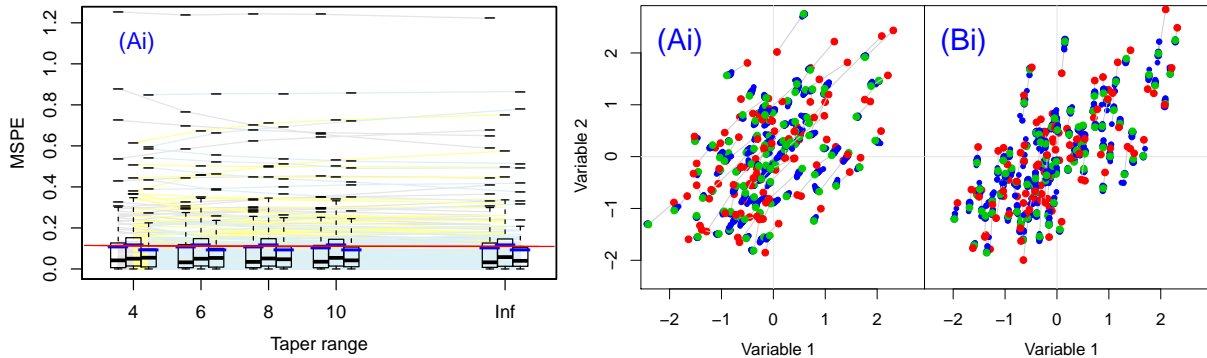


Figure 6: Left: Effect of increasing n on the prediction error. Horizontal red lines give the theoretical MSPEs. Within each boxplot triplet for a specific taper range, left is for $n = 400$ (gray), middle for 1024 (yellow), and right for 2500 (light blue). Prediction is based on $\hat{\theta}_{\text{tML}}$ with $\Delta = 1$ and 100 realizations of the bivariate process. Mean is indicated by the blue tick. Right: 100 bivariate predictions for $n = 400$ and $\Delta = 1$. Red: simulated “truth”, green: no tapering, blue: tapering with different taper ranges.

The right panel of Figure 6 shows the results of 100 bivariate predictions at the origin. There is again virtually no difference in the predictions using $\gamma = 4, 6, 8, 10$ (blue dots) and no tapering ($\gamma = \text{Inf}$, green dot). For smoother fields (variable 1, (B)), the prediction error is smaller and thus the difference between the red and blue/green dots is much smaller than for variable 2. The choice of the taper matrix function has again only a marginal effect on the result (not shown).

It has to be kept in mind that our simulation setup is the “least” favorable for the tapering approach. By including a nugget or reducing the spatial correlation we would receive even more appealing results because the importance of neighboring locations and their contribution to the prediction would be less important. Note also that estimation and prediction results can be improved by lowering Δ .

4.3 Computational efficiency

The analysis has been implemented with the freely available computer software R (Ihaka and Gentleman, 1996; R Development Core Team, 2015) running on a server with an Intel Xeon 6C E5-2640 2.50 GHz CPU (24 cores) and 256GB shared RAM (parallelization has not been explicitly exploited). The number of locations was kept below 2500 in order to maintain a reasonable computing time for the untapered settings, which require $\mathcal{O}(p^3n^3)$ computing time and $\mathcal{O}(p^2n^2)$ storage using straightforward R commands with classical methodologies.

The tapered settings have been implemented using sparse matrix data structures and algorithms. The package *spam* (Furrer, 2014; Furrer and Sain, 2010) is tailored in order to handle tapered covariance matrices, estimation, and prediction in the framework of Gaussian random fields. The core work load consists of calculating a Cholesky factorization of a permutation of the possibly tapered covariance matrix. The permutation (multiple minimum degree) improves storage and operation count; see Furrer and Sain (2010), Liu (1985), and Ng and Peyton (1993) for more technical details. From the Cholesky factor, it is straightforward to calculate the determinant as well as the quadratic term through two triangular solves. Hence, for large n , there is little difference in computational cost between a likelihood evaluation or a prediction. Exact operation counts are difficult to determine but the algorithms are virtually $\mathcal{O}(pnh^2)$ for operation count and $\mathcal{O}(pnh)$ for storage, where h is the “typical” number of observations within

the taper range.

For estimation, depending on the exact implementation, many likelihood evaluations are necessary. Using reasonable starting values, the R function `optim` required on average between 100 to 250 function evaluations depending on taper range and model ($n = 400$). In the untapered case, the average was typically somewhat lower. To reduce convergence issues, we started estimating the untapered version using the true parameter values as starting values and subsequently decreased the taper range using the previous optimum as starting values. Because of the large size of the datasets, no convergence issues were encountered and no sample was “manually” treated or eliminated.

5 Discussion and outlook

Similarly to the univariate case, multivariate tapering is a very effective approximation approach for prediction and for estimation of spatially correlated random processes. The small loss in prediction efficiency is recouped by the computational gains for reasonably large data sizes. For very large datasets, approximations have to be included and tapering is the method of choice as the computational implementation is straightforward. Compared with other approximation approaches (low-rank models, e.g., Cressie and Johannesson, 2008; Banerjee *et al.*, 2008; Stein, 2008, composite likelihood approaches, e.g., Stein *et al.*, 2004; Bevilacqua *et al.*, 2012; Eidsvik *et al.*, 2014, Gaussian Markov random fields type approximations, e.g., Hartman and Hössjer, 2008; Lindgren *et al.*, 2011, etc) tapering is the most accessible and most scalable approach.

Tapering is especially powerful for prediction. Even for very small tapers we have a MSPE that is almost identical to the MSPE for the untapered setting. However, we are substantially faster as a single prediction is roughly 20 and 100 times faster compared with a classical approach (for $n = 2500$ and $n = 10000$ using $\gamma = 5$). One likelihood evaluation is similarly computing intensive as a single prediction and thus the same advantages hold for estimation. If the ultimate goal is prediction, we advocate the use of the one-taper ML plugin estimates. The two-taper approach is computationally self-defeating and should only be used if unbiased estimates are absolutely necessary.

In the case where the different variables have a similar density of locations, we propose to use the same taper function for all direct and cross covariances. Compared with the taper range, the exact form of the taper plays a secondary role. Hence for different location sampling densities, possibly non-stationary, we foresee adaptive tapers as outlined by Anderes *et al.* (2013) or Bevilacqua *et al.* (2015) as a valuable alternative.

For estimation, the standard optimization routines of R (`optim` and its derivatives) require a substantial amount of time. We are currently experimenting with a simple grid search algorithm that would approximate the ML estimate sufficiently well. Based on the simulation results in the last section, if prediction based on plugin estimates is of interest, the approximation is sufficient.

While the uncertainty of the ML estimates can be harnessed through the Hessian (by product of the `optim` routine) sufficiently well, deriving uncertainty estimates for an entire prediction field remains a bottleneck, as accordingly many linear systems have to be solved.

Acknowledgments

RF acknowledges support of the UZH Research Priority Program (URPP) on “Global Change and Biodiversity” and the Swiss National Science Foundation SNSF-143282. FB presented the content of this paper at the statistics working group of the University of Vienna where he benefited from constructive comments.

Appendix

Proof of the theorems

Proof of Theorem 1. Because Θ is compact and because of Lemma 7, it is sufficient to show that, for any fixed θ , $L_\theta - \bar{L}_\theta = o_p(1)$. Hence, let an arbitrary θ be fixed. We have

$$\begin{aligned} L_\theta - \bar{L}_\theta &= \frac{1}{np} \log(\det[\Sigma_\theta \mathbf{K}_\theta^{-1}]) + \frac{1}{np} \mathbf{z}^\top (\Sigma_\theta^{-1} - \mathbf{K}_\theta^{-1}) \mathbf{z} \\ &= T_1 + T_2. \end{aligned} \tag{7}$$

We treat T_1 and T_2 separately. First

$$T_1 = \frac{1}{np} \sum_{i=1}^{np} \log\left(\lambda_i \left[\mathbf{K}_\theta^{-1/2} \Sigma_\theta \mathbf{K}_\theta^{-1/2}\right]\right).$$

The $\lambda_i(\cdot)$ above are between two constants $0 < A$ and $B < +\infty$ uniformly in i and n because of Condition 6 and Lemma 6. Thus, there exists a finite constant C so that for any i, n

$$\left| \log\left(\lambda_i \left[\mathbf{K}_\theta^{-1/2} \Sigma_\theta \mathbf{K}_\theta^{-1/2}\right]\right) \right| \leq C \left| 1 - \lambda_i \left[\mathbf{K}_\theta^{-1/2} \Sigma_\theta \mathbf{K}_\theta^{-1/2}\right] \right|.$$

Thus

$$\begin{aligned} |T_1| &\leq \frac{C}{np} \sum_{i=1}^{np} \left| 1 - \lambda_i \left[\mathbf{K}_\theta^{-1/2} \Sigma_\theta \mathbf{K}_\theta^{-1/2}\right] \right| \\ (\text{Cauchy-Schwarz:}) &\leq \frac{C}{np} \sqrt{np} \sqrt{\sum_{i=1}^{np} \left| 1 - \lambda_i \left[\mathbf{K}_\theta^{-1/2} \Sigma_\theta \mathbf{K}_\theta^{-1/2}\right] \right|^2} \\ &= C \sqrt{\frac{1}{np} \text{tr} \left(\left\{ \mathbf{I} - \mathbf{K}_\theta^{-1/2} \Sigma_\theta \mathbf{K}_\theta^{-1/2} \right\}^2 \right)} \\ &= C \sqrt{\frac{1}{np} \text{tr} \left(\left\{ \mathbf{K}_\theta^{-\frac{1}{2}} [\mathbf{K}_\theta - \Sigma_\theta] \mathbf{K}_\theta^{-\frac{1}{2}} \right\}^2 \right)} \\ &= C \sqrt{\frac{1}{np} \left\| \mathbf{K}_\theta^{-\frac{1}{2}} [\mathbf{K}_\theta - \Sigma_\theta] \mathbf{K}_\theta^{-\frac{1}{2}} \right\|_F^2}. \end{aligned}$$

Now, because of Condition 6, $\rho_1(\mathbf{K}_\theta^{-\frac{1}{2}})$ is bounded uniformly in n by a finite constant D . Hence we have

$$|T_1| \leq CD^2 \sqrt{\frac{1}{np} \|\mathbf{K}_\theta - \Sigma_\theta\|_F^2},$$

which goes to 0 as $n \rightarrow \infty$ because of Lemma 10. Next, turning to T_2 in (7),

$$\begin{aligned} \mathbb{E}(T_2) &= \frac{1}{np} \text{tr}(\Sigma_{\theta_0} (\Sigma_\theta^{-1} - \mathbf{K}_\theta^{-1})) \\ &= \frac{1}{np} \text{tr}(\Sigma_{\theta_0} \mathbf{K}_\theta^{-1} (\mathbf{K}_\theta - \Sigma_\theta) \Sigma_\theta^{-1}). \end{aligned}$$

Hence, interpreting $\text{tr}(\mathbf{A}\mathbf{B})$ as a scalar product between \mathbf{A} and \mathbf{B}^\top , we obtain by the Cauchy-Schwarz inequality

$$|\mathbb{E}(T_2)| \leq \sqrt{\frac{1}{np} \|\boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\Sigma}_{\theta_0} \mathbf{K}_\theta^{-1}\|_F^2} \sqrt{\frac{1}{np} \|\mathbf{K}_\theta - \boldsymbol{\Sigma}_\theta\|_F^2}.$$

In the above display, the first square root is bounded because of Condition 6 and of Lemma 6. The second square root goes to 0 because of Lemma 10. Hence $\mathbb{E}(T_2) \rightarrow_{n \rightarrow \infty} 0$. Furthermore

$$\begin{aligned} \text{Var}(T_2) &= \frac{2}{(np)^2} \text{tr}(\boldsymbol{\Sigma}_{\theta_0} [\boldsymbol{\Sigma}_\theta^{-1} - \mathbf{K}_\theta^{-1}] \boldsymbol{\Sigma}_{\theta_0} [\boldsymbol{\Sigma}_\theta^{-1} - \mathbf{K}_\theta^{-1}]) \\ &\leq \frac{2}{np} \rho_1(\boldsymbol{\Sigma}_{\theta_0})^2 [\rho_1(\boldsymbol{\Sigma}_\theta^{-1}) + \rho_1(\mathbf{K}_\theta^{-1})]^2. \end{aligned}$$

In the above display, the $\rho_1(\cdot)$ are bounded because of Condition 6 and Lemma 6. Thus $\text{Var}(T_2) \rightarrow_{n \rightarrow \infty} 0$. So $T_2 = o_p(1)$ which finishes the proof. \square

Proof of Theorem 3. We only prove (5), the proof of (4) being similar and technically simpler. Using $a^2 - b^2 = (a+b)(a-b)$ followed by the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} & \sup_{\boldsymbol{\theta}} \left| \int_{\mathcal{D}_n} [\boldsymbol{\sigma}_\theta(\mathbf{x})^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{z} - Z_1(\mathbf{x})]^2 f_n(\mathbf{x}) d\mathbf{x} - \int_{\mathcal{D}_n} [\mathbf{k}_\theta(\mathbf{x})^\top \mathbf{K}_\theta^{-1} \mathbf{z} - Z_1(\mathbf{x})]^2 f_n(\mathbf{x}) d\mathbf{x} \right| \\ & \leq \int_{\mathcal{D}_n} \sup_{\boldsymbol{\theta}} \left(\left| \boldsymbol{\sigma}_\theta(\mathbf{x})^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{z} + \mathbf{k}_\theta(\mathbf{x})^\top \mathbf{K}_\theta^{-1} \mathbf{z} - 2Z_1(\mathbf{x}) \right| \left| \boldsymbol{\sigma}_\theta(\mathbf{x})^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{z} - \mathbf{k}_\theta(\mathbf{x})^\top \mathbf{K}_\theta^{-1} \mathbf{z} \right| \right) f_n(\mathbf{x}) d\mathbf{x} \\ & \leq \sqrt{\int_{\mathcal{D}_n} \sup_{\boldsymbol{\theta}} (\boldsymbol{\sigma}_\theta(\mathbf{x})^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{z} + \mathbf{k}_\theta(\mathbf{x})^\top \mathbf{K}_\theta^{-1} \mathbf{z} - 2Z_1(\mathbf{x}))^2 f_n(\mathbf{x}) d\mathbf{x}} \\ & \quad \sqrt{\int_{\mathcal{D}_n} \sup_{\boldsymbol{\theta}} (\boldsymbol{\sigma}_\theta(\mathbf{x})^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{z} - \mathbf{k}_\theta(\mathbf{x})^\top \mathbf{K}_\theta^{-1} \mathbf{z})^2 f_n(\mathbf{x}) d\mathbf{x}} \\ & = \sqrt{U_1} \sqrt{U_2}. \end{aligned} \tag{8}$$

We show separately that $U_1 = O_p(1)$ and $U_2 = o_p(1)$. For U_1 ,

$$\begin{aligned} U_1 &\leq 3 \int_{\mathcal{D}_n} \sup_{\boldsymbol{\theta}} (\boldsymbol{\sigma}_\theta(\mathbf{x})^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{z})^2 f_n(\mathbf{x}) d\mathbf{x} + 3 \int_{\mathcal{D}_n} \sup_{\boldsymbol{\theta}} (\mathbf{k}_\theta(\mathbf{x})^\top \mathbf{K}_\theta^{-1} \mathbf{z})^2 f_n(\mathbf{x}) d\mathbf{x} \\ &\quad + 12 \int_{\mathcal{D}_n} \sup_{\boldsymbol{\theta}} (Z_1(\mathbf{x}))^2 f_n(\mathbf{x}) d\mathbf{x}. \end{aligned}$$

The last random integral in the above display has constant mean value $12c_{11}(\mathbf{0}; \boldsymbol{\theta}_0)$ so it is bounded in probability. We address the two remaining random integrals in the same way, and give the details for the first one only. Using a version of Sobolev embedding theorem (Theorem 4.12, Part I, Case A in Adams and Fournier, 2003), there exists a finite constant A_Θ depending only on Θ so that

$$\sup_{\boldsymbol{\theta}} (\boldsymbol{\sigma}_\theta(\mathbf{x})^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{z})^2 \leq A_\Theta \int_{\Theta} \left| (\boldsymbol{\sigma}_\theta(\mathbf{x})^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{z})^2 \right|^{q+1} d\boldsymbol{\theta} + A_\Theta \sum_{i=1}^q \int_{\Theta} \left| \frac{\partial}{\partial \theta_i} \left[(\boldsymbol{\sigma}_\theta(\mathbf{x})^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{z})^2 \right] \right|^{q+1} d\boldsymbol{\theta}.$$

Hence, using Fubini theorem for non-negative integrand and $(|a| + |b|)^{q+1} \leq 2^{q+1}(|a|^{q+1} + |b|^{q+1})$,

we obtain

$$\begin{aligned}
& \mathbb{E} \left(\int_{\mathcal{D}_n} \sup_{\boldsymbol{\theta}} \left(\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x})^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} \right)^2 f_n(\mathbf{x}) d\mathbf{x} \right) \\
& \leq A_{\Theta} \int_{\Theta} \int_{\mathcal{D}_n} \mathbb{E} \left(\left| \left(\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x})^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} \right)^2 \right|^{q+1} \right) f_n(\mathbf{x}) d\mathbf{x} d\boldsymbol{\theta} \\
& \quad + A_{\Theta} 2^{2q+2} \sum_{i=1}^q \int_{\Theta} \int_{\mathcal{D}_n} \mathbb{E} \left(\left| \left(\frac{\partial \boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x})^\top}{\partial \theta_i} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} \right) \left(\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x})^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} \right) \right|^{q+1} \right) f_n(\mathbf{x}) d\mathbf{x} d\boldsymbol{\theta} \\
& \quad + A_{\Theta} 2^{2q+2} \sum_{i=1}^q \int_{\Theta} \int_{\mathcal{D}_n} \mathbb{E} \left(\left| \left(\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x})^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} \right) \left(\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x})^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} \right) \right|^{q+1} \right) f_n(\mathbf{x}) d\mathbf{x} d\boldsymbol{\theta}.
\end{aligned}$$

Let $\lambda(\Theta)$ be the Lebesgue measure of Θ . Using the Cauchy–Schwarz inequality and letting B_{q+1} be the positive constant so that, for X following a Gaussian distribution with zero mean, $\mathbb{E}(X^{2(q+1)}) = B_{q+1}(\mathbb{E}(X^2))^{q+1}$, we obtain, by letting $D = A_{\Theta} B_{q+1} \lambda(\Theta) 2^{2q+2}$,

$$\begin{aligned}
& \mathbb{E} \left(\int_{\mathcal{D}_n} \sup_{\boldsymbol{\theta}} \left(\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x})^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} \right)^2 f_n(\mathbf{x}) d\mathbf{x} \right) \tag{9} \\
& \leq A_{\Theta} B_{q+1} \lambda(\Theta) \sup_{\mathbf{x}, \boldsymbol{\theta}} \mathbb{E}^{q+1} \left(\left(\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x})^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} \right)^2 \right) \\
& \quad + D \sum_{i=1}^q \sup_{\mathbf{x}, \boldsymbol{\theta}} \sqrt{\mathbb{E}^{q+1} \left(\left(\frac{\partial \boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x})^\top}{\partial \theta_i} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} \right)^2 \right)} \sup_{\mathbf{x}, \boldsymbol{\theta}} \sqrt{\mathbb{E}^{q+1} \left(\left(\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x})^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} \right)^2 \right)} \\
& \quad + D \sum_{i=1}^q \sup_{\mathbf{x}, \boldsymbol{\theta}} \sqrt{\mathbb{E}^{q+1} \left(\left(\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x})^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} \right)^2 \right)} \sup_{\mathbf{x}, \boldsymbol{\theta}} \sqrt{\mathbb{E}^{q+1} \left(\left(\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x})^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} \right)^2 \right)}.
\end{aligned}$$

Now, all the $\mathbb{E}^{q+1}(\cdot)$ above are of the form $\mathbb{E}^{q+1}([\mathbf{w}_{\boldsymbol{\theta}}(\mathbf{x})^\top \mathbf{M}_{\boldsymbol{\theta}} \mathbf{z}]^2)$. Furthermore, $\mathbf{M}_{\boldsymbol{\theta}}$ is symmetric and satisfies, by using Condition 6 and Lemma 6, $\sup_{\boldsymbol{\theta}} \rho_1(\mathbf{M}_{\boldsymbol{\theta}}) \leq C$ for a finite constant C . Finally, for $i = k(n-1) + a$, with $k = 1, \dots, p$ and $a = 1, \dots, n$, $\sup_{\boldsymbol{\theta}} |\mathbf{w}_{\boldsymbol{\theta}}(\mathbf{x})_i| \leq G/(1 + |\mathbf{x} - \mathbf{x}_a|^{d+\alpha})$, for a finite constant G . Hence,

$$\begin{aligned}
\sup_{\mathbf{x}, \boldsymbol{\theta}} \mathbb{E}([\mathbf{w}_{\boldsymbol{\theta}}(\mathbf{x})^\top \mathbf{M}_{\boldsymbol{\theta}} \mathbf{z}]^2) &= \sup_{\mathbf{x}, \boldsymbol{\theta}} \mathbf{w}_{\boldsymbol{\theta}}(\mathbf{x})^\top \mathbf{M}_{\boldsymbol{\theta}} \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0} \mathbf{M}_{\boldsymbol{\theta}} \mathbf{w}_{\boldsymbol{\theta}}(\mathbf{x}) \\
&\leq \sup_{\mathbf{x}, \boldsymbol{\theta}} \|\mathbf{w}_{\boldsymbol{\theta}}(\mathbf{x})\|^2 C^2 \sup_{\boldsymbol{\theta}} \rho_1(\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}),
\end{aligned}$$

which is bounded because of Lemmas 4 and 6. Hence, in (8), $U_1 = O_p(1)$. Let us now turn to U_2 . Using the Sobolev embedding theorem again with the constant A_{Θ} , we obtain

$$\begin{aligned}
\mathbb{E}(U_2) &\leq A_{\Theta} \int_{\Theta} \int_{\mathcal{D}_n} \mathbb{E} \left(\left| \left[\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x})^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} - \mathbf{k}_{\boldsymbol{\theta}}(\mathbf{x})^\top \mathbf{K}_{\boldsymbol{\theta}}^{-1} \mathbf{z} \right]^2 \right|^{q+1} \right) f_n(\mathbf{x}) d\mathbf{x} d\boldsymbol{\theta} \\
&\quad + A_{\Theta} \sum_{i=1}^q \int_{\Theta} \int_{\mathcal{D}_n} \mathbb{E} \left(\left| \frac{\partial}{\partial \theta_i} \left(\left[\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{x})^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} - \mathbf{k}_{\boldsymbol{\theta}}(\mathbf{x})^\top \mathbf{K}_{\boldsymbol{\theta}}^{-1} \mathbf{z} \right]^2 \right) \right|^{q+1} \right) f_n(\mathbf{x}) d\mathbf{x} d\boldsymbol{\theta} \\
&= A_{\Theta} I_0 + A_{\Theta} \sum_{i=1}^q I_i
\end{aligned}$$

In the above display, we only show that the integrals I_1, \dots, I_q converge to 0, since it is more difficult than for the integral I_0 . Hence let us fix an integer i in $\{1, \dots, q\}$. Using Cauchy–Schwarz

inequality, we have

$$I_i \leq A_\Theta \lambda(\Theta) 2^{q+1} \sup_{\mathbf{x}, \boldsymbol{\theta}} \sqrt{\mathbb{E} \left(\left| \frac{\partial}{\partial \theta_i} [\boldsymbol{\sigma}_\theta(\mathbf{x})^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{z} - \mathbf{k}_\theta(\mathbf{x})^\top \mathbf{K}_\theta^{-1} \mathbf{z}] \right|^{2(q+1)} \right)} \\ \times \sup_{\mathbf{x}, \boldsymbol{\theta}} \sqrt{\mathbb{E} \left(\left| \boldsymbol{\sigma}_\theta(\mathbf{x})^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{z} - \mathbf{k}_\theta(\mathbf{x})^\top \mathbf{K}_\theta^{-1} \mathbf{z} \right|^{2(q+1)} \right)}.$$

Again, both of the supremums of square roots in the above display go to 0 as $n \rightarrow \infty$ and we show it only for the first one, since it is more difficult than for the second one. Using the positive constant B_{q+1} used before (9), it is sufficient to show that

$$\sup_{\boldsymbol{\theta}, \mathbf{x}} \mathbb{E} \left(\left\{ \frac{\partial}{\partial \theta_i} [\boldsymbol{\sigma}_\theta(\mathbf{x})^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{z} - \mathbf{k}_\theta(\mathbf{x})^\top \mathbf{K}_\theta^{-1} \mathbf{z}] \right\}^2 \right)$$

goes to 0 as $n \rightarrow \infty$. Then, we use

$$(a_{11} - a_{22})^2 \leq 2 [(a_{11} - a_{21})^2 + (a_{21} - a_{22})^2]$$

and

$$(b_{1111} - b_{2222})^2 \leq 4 [(b_{1111} - b_{2111})^2 + (b_{2111} - b_{2211})^2 + (b_{2211} - b_{2221})^2 + (b_{2221} - b_{2222})^2],$$

where subscripts 1 and 2 denote ‘‘untapered’’ and ‘‘tapered’’ and where for example $a_{21} = \{[\partial \mathbf{k}_\theta(\mathbf{x})]/[\partial \theta_i]\}^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{z}$ and $b_{2211} = \mathbf{k}_\theta(\mathbf{x})^\top \mathbf{K}_\theta^{-1} \{[\partial \boldsymbol{\Sigma}_\theta]/[\partial \theta_i]\} \boldsymbol{\Sigma}_\theta^{-1} \mathbf{z}$. From this, it is sufficient to show that a generic term of the form

$$\sup_{\boldsymbol{\theta}, \mathbf{x}} \mathbb{E} \left(\left[(\mathbf{v}_\theta(\mathbf{x}) - \mathbf{w}_\theta(\mathbf{x}))^\top \mathbf{M}_\theta \mathbf{z} \right]^2 \right), \quad (10)$$

$$\sup_{\boldsymbol{\theta}, \mathbf{x}} \mathbb{E} \left(\left[\mathbf{m}_\theta(\mathbf{x})^\top \mathbf{M}_\theta (\boldsymbol{\Sigma}_\theta^{-1} - \mathbf{K}_\theta^{-1}) \mathbf{N}_\theta \mathbf{z} \right]^2 \right) \quad (11)$$

or

$$\sup_{\boldsymbol{\theta}, \mathbf{x}} \mathbb{E} \left(\left[\mathbf{m}_\theta(\mathbf{x})^\top \mathbf{M}_\theta \left(\frac{\partial \boldsymbol{\Sigma}_\theta}{\partial \theta_i} - \frac{\partial \mathbf{K}_\theta}{\partial \theta_i} \right) \mathbf{N}_\theta \mathbf{z} \right]^2 \right), \quad (12)$$

goes to 0. In (10), (11) and (12), $\sup_{\boldsymbol{\theta}} \rho_1(\mathbf{M}_\theta)$ and $\sup_{\boldsymbol{\theta}} \rho_1(\mathbf{N}_\theta)$ are bounded (Condition 6 and Lemma 6); $\mathbf{v}_\theta(\mathbf{x}) - \mathbf{w}_\theta(\mathbf{x}) = \boldsymbol{\sigma}_\theta(\mathbf{x}) - \mathbf{k}_\theta(\mathbf{x})$ or $\mathbf{v}_\theta(\mathbf{x}) - \mathbf{w}_\theta(\mathbf{x}) = (\partial \boldsymbol{\sigma}_\theta(\mathbf{x})) / (\partial \theta_i) - (\partial \mathbf{k}_\theta(\mathbf{x})) / (\partial \theta_i)$; and $\mathbf{m}_\theta(\mathbf{x}) = \mathbf{k}_\theta(\mathbf{x})$ or $\mathbf{m}_\theta(\mathbf{x}) = \{[\partial \mathbf{k}_\theta(\mathbf{x})]/[\partial \theta_i]\}$.

Let us now show that a generic term of the form (10) goes to 0. We have

$$\sup_{\boldsymbol{\theta}, \mathbf{x}} \mathbb{E} \left(\left[(\mathbf{v}_\theta(\mathbf{x}) - \mathbf{w}_\theta(\mathbf{x}))^\top \mathbf{M}_\theta \mathbf{z} \right]^2 \right) = \sup_{\boldsymbol{\theta}, \mathbf{x}} (\mathbf{v}_\theta(\mathbf{x}) - \mathbf{w}_\theta(\mathbf{x}))^\top \mathbf{M}_\theta \boldsymbol{\Sigma}_{\theta_0} \mathbf{M}_\theta^\top (\mathbf{v}_\theta(\mathbf{x}) - \mathbf{w}_\theta(\mathbf{x})) \\ \leq \sup_{\boldsymbol{\theta}} \rho_1(\mathbf{M}_\theta \boldsymbol{\Sigma}_{\theta_0} \mathbf{M}_\theta^\top) \sup_{\boldsymbol{\theta}, \mathbf{x}} \|\mathbf{v}_\theta(\mathbf{x}) - \mathbf{w}_\theta(\mathbf{x})\|^2,$$

which goes to 0 as $n \rightarrow \infty$ by remembering that $\sup_{\boldsymbol{\theta}} \rho_1(\mathbf{M}_\theta)$ is bounded and by using Lemmas 6 and 8.

For a generic term of the form (11), we have

$$\begin{aligned}
& \sup_{\boldsymbol{\theta}, \mathbf{x}} \mathbb{E} \left(\left[\mathbf{m}_{\boldsymbol{\theta}}(\mathbf{x})^\top \mathbf{M}_{\boldsymbol{\theta}} (\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} - \mathbf{K}_{\boldsymbol{\theta}}^{-1}) \mathbf{N}_{\boldsymbol{\theta}} \mathbf{z} \right]^2 \right) \\
&= \sup_{\boldsymbol{\theta}, \mathbf{x}} \mathbb{E} \left(\left[\mathbf{m}_{\boldsymbol{\theta}}(\mathbf{x})^\top \mathbf{M}_{\boldsymbol{\theta}} \mathbf{K}_{\boldsymbol{\theta}}^{-1} (\mathbf{K}_{\boldsymbol{\theta}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{N}_{\boldsymbol{\theta}} \mathbf{z} \right]^2 \right) \tag{13} \\
&= \sup_{\boldsymbol{\theta}, \mathbf{x}} \mathbf{m}_{\boldsymbol{\theta}}(\mathbf{x})^\top \mathbf{M}_{\boldsymbol{\theta}} \mathbf{K}_{\boldsymbol{\theta}}^{-1} (\mathbf{K}_{\boldsymbol{\theta}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{N}_{\boldsymbol{\theta}} \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0} \mathbf{N}_{\boldsymbol{\theta}}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} (\mathbf{K}_{\boldsymbol{\theta}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \mathbf{K}_{\boldsymbol{\theta}}^{-1} \mathbf{M}_{\boldsymbol{\theta}}^\top \mathbf{m}_{\boldsymbol{\theta}}(\mathbf{x}) \\
&\leq \sup_{\boldsymbol{\theta}, \mathbf{x}} \|\mathbf{m}_{\boldsymbol{\theta}}(\mathbf{x})\|^2 \rho_1(\mathbf{M}_{\boldsymbol{\theta}})^2 \rho_1(\mathbf{N}_{\boldsymbol{\theta}})^2 \rho_1(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1})^2 \rho_1(\mathbf{K}_{\boldsymbol{\theta}}^{-1})^2 \rho_1(\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}) \rho_1(\mathbf{K}_{\boldsymbol{\theta}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}})^2.
\end{aligned}$$

In the above display, $\sup_{\boldsymbol{\theta}, \mathbf{x}} \|\mathbf{m}_{\boldsymbol{\theta}}(\mathbf{x})\|^2$ is bounded because of Lemma 4. Furthermore all the $\rho_1(\cdot)^2$, except the last one are bounded uniformly in $\boldsymbol{\theta}$, by remembering that $\sup_{\boldsymbol{\theta}} \rho_1(\mathbf{M}_{\boldsymbol{\theta}})$ and $\sup_{\boldsymbol{\theta}} \rho_1(\mathbf{N}_{\boldsymbol{\theta}})$ are bounded, and because of Condition 6 and Lemma 6. Finally $\sup_{\boldsymbol{\theta}} \rho_1(\mathbf{K}_{\boldsymbol{\theta}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ goes to 0 as $n \rightarrow \infty$ because of Lemma 9. Hence a generic term of the form (11) goes to 0 as $n \rightarrow \infty$. Finally, by the same arguments as following (13), we show that a generic term of the form (12) goes to 0 as $n \rightarrow \infty$. Hence, $\mathbb{E}(U_2)$ in (8) goes to 0 as $n \rightarrow \infty$ which concludes the proof. \square

Technical results

The following lemma is a generalization of Lemma D.1 in Bachoc (2014b).

Lemma 4. *Let $\Delta > 0$ and $\alpha > 0$ be fixed. Let $f(\mathbf{x}; \boldsymbol{\theta})$ be a family of functions: $\mathbb{R}^d \rightarrow \mathbb{R}$ so that for all $\boldsymbol{\theta} \in \Theta$, $|f(\mathbf{x}; \boldsymbol{\theta})| \leq 1/(1 + |\mathbf{x}|^{d+\alpha})$. Then, for any $m \in \mathbb{N}^+$, $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{s}_1, \dots, \mathbf{s}_m \in \mathbb{R}^d$, so that for any $i \neq j$ $|\mathbf{s}_i - \mathbf{s}_j| \geq \Delta$, we have*

$$\sup_{\boldsymbol{\theta}} \sum_{i=1}^m |f(\mathbf{s}_i - \mathbf{v}; \boldsymbol{\theta})| \leq \frac{d2^{2d}}{\Delta^d} \sum_{k=1}^{+\infty} \frac{k^{d-1}}{1 + (k-1)^{d+\alpha}},$$

where the right-hand term in the above display is a finite constant depending only on d , Δ and α .

Proof of Lemma 4. By assumption on $f(\mathbf{x}, \boldsymbol{\theta})$ we have

$$\sup_{\boldsymbol{\theta}} \sum_{i=1}^m |f(\mathbf{s}_i - \mathbf{v}; \boldsymbol{\theta})| \leq \sum_{i=1}^m \frac{1}{1 + |\mathbf{s}_i - \mathbf{v}|^{d+\alpha}}.$$

Let, for $k \geq 1$, N_k be the number of points \mathbf{s}_j in $E_k = \{\mathbf{w}; |\mathbf{w} - \mathbf{v}| \leq k\} \setminus \{\mathbf{w}; |\mathbf{w} - \mathbf{v}| \leq k-1\}$. Then, to the N_k points \mathbf{s}_j that are in E_k we can associate N_k disjoint $|\cdot|$ -balls in E_k so that each of them has volume $(\Delta/2)^d$ (recall $|\mathbf{a}| = \max_l |a_l|$). The total volume occupied by these balls is $N_k(\Delta/2)^d$. On the other hand, the volume of E_k is

$$(2k)^d - (2k-2)^d = 2^d \int_{k-1}^k du^{d-1} du \leq 2^d dk^{d-1}.$$

So we have $N_k \leq d2^{2d}k^{d-1}/\Delta^d$. The result is then obtained by noting that for $\mathbf{s}_j \in E_k$, $|\mathbf{s}_j - \mathbf{v}| \geq k-1$. \square

The following lemma is a generalization of Lemma D.3 in Bachoc (2014b).

Lemma 5. Consider the setting of Lemma 4. Then, for any $N \in \mathbb{N}^+$, for any $m \in \mathbb{N}^+$, $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{s}_1, \dots, \mathbf{s}_m \in \mathbb{R}^d$, so that for any $i \neq j$ $|\mathbf{s}_i - \mathbf{s}_j| \geq \Delta$, we have

$$\sup_{\boldsymbol{\theta}} \sum_{i=1, \dots, m; |\mathbf{s}_i - \mathbf{v}| > N-1} |f(\mathbf{s}_i - \mathbf{v}; \boldsymbol{\theta})| \leq \frac{d2^{2d}}{\Delta^d} \sum_{k=N}^{+\infty} \frac{k^{d-1}}{1 + (k-1)^{d+\alpha}},$$

where the right-hand term in the above display is a function of N , d , Δ and α only, that goes to 0 as $N \rightarrow +\infty$ and for fixed d, Δ, α .

Proof of Lemma 5. The lemma is obtained by the proof of Lemma 4, by noting that only the points \mathbf{s}_j that are in E_k for $k \geq N$ give a non-zero contribution to the sum in the left-hand side of the display in the lemma. \square

Lemma 6. Assume that Condition 5 holds. Let $f_{kl}(\mathbf{x}; \boldsymbol{\theta})$, $k, l = 1, \dots, p$ be p^2 functions: $\mathbb{R}^d \rightarrow \mathbb{R}$ so that for all $\boldsymbol{\theta} \in \Theta$, $|f_{kl}(\mathbf{x}; \boldsymbol{\theta})| \leq 1/(1+|\mathbf{x}|^{d+\alpha})$ and $f_{kl}(\mathbf{x}; \boldsymbol{\theta}) = f_{lk}(-\mathbf{x}; \boldsymbol{\theta})$. Let $\mathbf{F}_{\boldsymbol{\theta}}$ be the $np \times np$ matrix defined by, for $i = (k-1)n+a$ and $j = (l-1)n+b$, with $k, l = 1, \dots, p$ and $a, b = 1, \dots, n$, $f_{\boldsymbol{\theta}ij} = f_{kl}(\mathbf{x}_a - \mathbf{x}_b; \boldsymbol{\theta})$. Then, there exists a constant $A < \infty$ so that for any n , $\boldsymbol{\theta}$, $\rho_1(\mathbf{F}_{\boldsymbol{\theta}}) \leq A$.

Proof of Lemma 6. Since $\mathbf{F}_{\boldsymbol{\theta}}$ is symmetric, $\rho_1(\mathbf{F}_{\boldsymbol{\theta}}) = \lambda_1(\mathbf{F}_{\boldsymbol{\theta}})$. Hence, because of Gershgorin circle theorem and of $|f_{\boldsymbol{\theta}kk}| \leq 1$ for any $n, \boldsymbol{\theta}$, it is sufficient to show that

$$\sup_{i, n, \boldsymbol{\theta}} \sum_{j=1, \dots, np; j \neq i} |f_{\boldsymbol{\theta}ij}|$$

is finite. By writing the sum above as the sum of p subsums, it is sufficient to show that

$$\sup_{k, l, a, n, \boldsymbol{\theta}} \sum_{j=1, \dots, n} |f_{kl}(\mathbf{x}_a - \mathbf{x}_j; \boldsymbol{\theta})|$$

is finite. This is true because of Lemma 4. \square

Lemma 7. Assume that conditions 3, 5, and 6 hold. Then, as $n \rightarrow \infty$

$$\sup_{i, \boldsymbol{\theta}} \left| \frac{\partial}{\partial \theta_i} L_{\boldsymbol{\theta}} \right| = O_p(1) \quad \text{and} \quad \sup_{i, \boldsymbol{\theta}} \left| \frac{\partial}{\partial \theta_i} \bar{L}_{\boldsymbol{\theta}} \right| = O_p(1).$$

Proof of Lemma 7. We do the proof for $L_{\boldsymbol{\theta}}$ only since the proof for $\bar{L}_{\boldsymbol{\theta}}$ is identical. We have for any $i = 1, \dots, q$,

$$\begin{aligned} \sup_{\boldsymbol{\theta} \in \Theta} \left| \frac{\partial}{\partial \theta_i} L_{\boldsymbol{\theta}} \right| &= \sup_{\boldsymbol{\theta} \in \Theta} \left| \frac{1}{np} \text{tr} \left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i} \right) - \frac{1}{np} \mathbf{z}^{\top} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{z} \right| \\ &\leq \sup_{\boldsymbol{\theta}} \rho_1 \left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1/2} \frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1/2} \right) + \frac{1}{np} \mathbf{z}^{\top} \mathbf{z} \sup_{\boldsymbol{\theta}} \rho_1 \left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \right). \end{aligned}$$

Now, $(1/(np))\mathbf{z}^{\top} \mathbf{z}$ is bounded in probability since it is positive with constant mean value $(1/p) \sum_{k=1}^p c_{kk}(\mathbf{0}; \boldsymbol{\theta}_0)$. The two $\rho_1(\cdot)$ in the above display are bounded uniformly in $\boldsymbol{\theta}$ because of $\rho_1(\mathbf{CD}) \leq \rho_1(\mathbf{C})\rho_1(\mathbf{D})$, of Conditions 3, 5, and 6 and of Lemma 6. \square

Lemma 8. Let $\alpha > 0$ and $\Delta > 0$ be fixed. Let $f(\mathbf{x}; \boldsymbol{\theta})$ be a family of functions: $\mathbb{R}^d \rightarrow \mathbb{R}$ so that for all $\boldsymbol{\theta}$, $|f(\mathbf{x}; \boldsymbol{\theta})| \leq 1/(1+|\mathbf{x}|^{d+\alpha})$. Let $t(\mathbf{x})$ be a fixed function: $\mathbb{R}^d \rightarrow \mathbb{R}$ that is continuous at $\mathbf{0}$ and so that $t(\mathbf{0}) = 1$ and $|t(\mathbf{x})| \leq 1$. Let S_m be the set of all sets of points $(\mathbf{s}_1, \dots, \mathbf{s}_m)$ so that for $i \neq j$ $|\mathbf{s}_i - \mathbf{s}_j| \geq \Delta$. Then,

$$\sup_{m, (\mathbf{s}_1, \dots, \mathbf{s}_m) \in S_m, \mathbf{v}, \boldsymbol{\theta}} \sum_{i=1}^m |f(\mathbf{v} - \mathbf{s}_i; \boldsymbol{\theta}) - f(\mathbf{v} - \mathbf{s}_i; \boldsymbol{\theta})t((\mathbf{v} - \mathbf{s}_i)/\gamma)|$$

goes to 0 as $\gamma \rightarrow \infty$.

Proof of Lemma 8. Let $\epsilon > 0$ be fixed. Because of Lemma 5, we can find $M \in \mathbb{N}^+$ so that

$$\sup_{m, (\mathbf{s}_1, \dots, \mathbf{s}_m) \in S_m, \mathbf{v}, \boldsymbol{\theta}} \sum_{i=1, \dots, m; |\mathbf{v} - \mathbf{s}_i| > M-1} |f(\mathbf{v} - \mathbf{s}_i; \boldsymbol{\theta}) - f(\mathbf{v} - \mathbf{s}_i; \boldsymbol{\theta})t((\mathbf{v} - \mathbf{s}_i)/\gamma)| \leq \epsilon.$$

Because t is continuous at $\mathbf{0}$, we have for γ large enough and for $|\mathbf{v} - \mathbf{s}_i| \leq M - 1$

$$|1 - t((\mathbf{s}_i - \mathbf{v})/\gamma)| \leq \frac{\epsilon}{\tilde{N}_{M-1}},$$

where \tilde{N}_{M-1} is the maximum numbers of points \mathbf{s}_j so that $|\mathbf{s}_j - \mathbf{v}| \leq M - 1$, over all possible m , \mathbf{v} and $(\mathbf{s}_1, \dots, \mathbf{s}_m) \in S_m$. Putting the two bounds together, and using $|f(\mathbf{x}; \boldsymbol{\theta})| \leq 1$ we obtain, for γ large enough,

$$\sup_{m, (\mathbf{s}_1, \dots, \mathbf{s}_m) \in S_m, \boldsymbol{\theta}} \sum_{i=1}^m |f(\mathbf{v} - \mathbf{s}_i; \boldsymbol{\theta}) - f(\mathbf{v} - \mathbf{s}_i; \boldsymbol{\theta})t((\mathbf{v} - \mathbf{s}_i)/\gamma)| \leq \epsilon + \tilde{N}_{M-1} \frac{\epsilon}{\tilde{N}_{M-1}},$$

which finishes the proof. \square

Lemma 9. *Assume that Conditions 4 and 5 hold. Let $f_{kl}(\mathbf{x}; \boldsymbol{\theta})$ and $\mathbf{F}_{\boldsymbol{\theta}}$ be as in Lemma 6. Let $t_{kl}(\mathbf{x})$, $k, l = 1, \dots, p$, be the p^2 taper functions satisfying Condition 4. Let γ be the taper range, also satisfying Condition 4. Let $\mathbf{G}_{\boldsymbol{\theta}}$ be the $np \times np$ matrix defined by, for $i = (k - 1)n + a$ and $j = (l - 1)n + b$, with $k, l = 1, \dots, p$ and $a, b = 1, \dots, n$, $g_{\boldsymbol{\theta}ij} = f_{kl}(\mathbf{x}_a - \mathbf{x}_b; \boldsymbol{\theta})t_{kl}((\mathbf{x}_a - \mathbf{x}_b)/\gamma)$. Then, $\sup_{\boldsymbol{\theta}} \rho_1(\mathbf{F}_{\boldsymbol{\theta}} - \mathbf{G}_{\boldsymbol{\theta}}) \rightarrow_{n \rightarrow \infty} 0$.*

Proof of Lemma 9. The lemma is a consequence of Lemma 8. The proof is based on Gershgorin circle theorem as for the proof of Lemma 6. \square

Lemma 10. *Assume that Conditions 3, 4, and 5 hold. Then, $\sup_{\boldsymbol{\theta}} \frac{1}{np} \|\boldsymbol{\Sigma}_{\boldsymbol{\theta}} - \mathbf{K}_{\boldsymbol{\theta}}\|_F^2$ goes to 0 as $n \rightarrow \infty$.*

Proof of Lemma 10. The lemma is a consequence of Lemma 9. \square

References

- Abramowitz, M. and Stegun, I. A., editors (1970). *Handbook of Mathematical Functions*. Dover, New York.
- Adams, R. A. and Fournier, J. J. F. (2003). *Sobolev Spaces*. Academic Press, Amsterdam.
- Anderes, E., Huser, R., Nychka, D., and Coram, M. (2013). Nonstationary positive definite tapering on the plane. *J. Comput. Graph. Stat.*, **22**, 848–865.
- Bachoc, F. (2014a). Asymptotic analysis of covariance parameter estimation for gaussian processes in the misspecified case. arXiv preprint <http://arxiv.org/abs/1412.1926>.
- Bachoc, F. (2014b). Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of gaussian processes. *J. Multivariate Anal.*, **125**, 1–35.
- Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Statist. Soc. B.*, **70**, 825–848.

- Bevilacqua, M., Gaetan, C., Mateu, J., and Porcu, E. (2012). Estimating space and space-time covariance functions for large data sets: A weighted composite likelihood approach. *J. Amer. Statist. Assoc.*, **107**, 268–280.
- Bevilacqua, M., Genton, M., Porcu, E., and Zastavnyi, V. (2015). Adaptive tapering for space-time covariance functions. Submitted.
- Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc. B*, **70**, 209–226.
- Daley, D. J., Porcu, E., and Bevilacqua, M. (2014). Classes of compactly supported covariance functions for multivariate random fields. *Stoch. Environ. Res. Risk Assess.*, **29**, 1–15.
- Demel, S. S. (2013). *Modeling and computations of multivariate datasets in space and time*. PhD thesis, Kansas State University, Manhattan, Kansas.
- Du, J., Zhang, H., and Mandrekar, V. S. (2009). Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *Ann. Statist.*, **37**, 3330–3361.
- Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J. (2014). Estimation and prediction in spatial models with block composite likelihoods. *J. Comput. Graph. Stat.*, **23**, 295–315.
- Furrer, R. (2014). *spam: SPArse Matrix*. R package version 1.0-1, <http://cran.r-project.org/web/packages/spam>.
- Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Stat.*, **15**, 502–523.
- Furrer, R. and Sain, S. R. (2010). spam: A sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields. *Journal of Statistical Software*, **36**, 1–25.
- Gneiting, T., Kleiber, W., and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *J. Amer. Statist. Assoc.*, **105**, 1167–1177.
- Hartman, L. and Hössjer, O. (2008). Fast kriging of large data sets with Gaussian Markov random fields. *Comput. Stat. Data An.*, **52**, 2331–2349.
- Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.
- Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets. *J. Amer. Statist. Assoc.*, **103**, 1545–1555.
- Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Statist. Soc. B*, **73**, 423–498.
- Liu, J. W. H. (1985). Modification of the minimum-degree algorithm by multiple elimination. *ACM Trans. Math. Softw.*, **11**, 141–153.
- Ma, C. (2011a). Covariance matrices for second-order vector random fields in space and time. *IEEE Trans. Signal Process.*, **59**, 2160–2168.

- Ma, C. (2011b). Vector random fields with long range dependence. *Fractals*, **19**, 249–258.
- Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**, 135–146.
- Ng, E. G. and Peyton, B. W. (1993). Block sparse Cholesky algorithms on advanced uniprocessor computers. *SIAM J. Sci. Comput.*, **14**, 1034–1056.
- R Development Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>.
- Ruiz-Medina, M. D. and Porcu, E. (2015). Equivalence of gaussian measures of multivariate random fields. *Stoch. Environ. Res. Risk Assess.*, **29**, 325–334.
- Sain, S. R., Furrer, R., and Cressie, N. (2011). A spatial analysis of multivariate output from regional climate models. *Ann. Appl. Stat.*, **5**, 150–175.
- Shaby, B. A. and Ruppert, D. (2012). Tapered covariance: Bayesian estimation and asymptotics. *J. Comput. Graph. Stat.*, **21**, 433–452.
- Stein, M. L. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *Ann. Statist.*, **16**, 55–63.
- Stein, M. L. (1990). Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. *Ann. Statist.*, **18**, 850–872.
- Stein, M. L. (1997). Efficiency of linear predictors for periodic processes using an incorrect covariance function. *J. Statist. Plann. Inference*, **58**, 321–331.
- Stein, M. L. (1999). Predicting random fields with increasing dense observations. *Ann. Appl. Probab.*, **9**, 242–273.
- Stein, M. L. (2002). The screening effect in kriging. *Ann. Statist.*, **30**, 298–323.
- Stein, M. L. (2008). A modeling approach for large spatial datasets. *J. Korean Stat. Soc.*, **37**, 3–10.
- Stein, M. L. (2013). Statistical properties of covariance tapers. *J. Comput. Graph. Stat.*, **22**, 866–885.
- Stein, M. L., Chi, Z., and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *J. R. Statist. Soc. B*, **66**, 275–296.
- Wang, D. and Loh, W.-L. (2011). On fixed-domain asymptotics and covariance tapering in gaussian random field models. *Electron. J. Statist.*, **5**, 238–269.
- Watkins, A. and Al-Boutiahi, F. (1990). On maximum likelihood estimation of parameters in incorrectly specified models of covariance for spatial data. *Math. Geol.*, **22**, 151–173.
- Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.*, **4**, 389–396.