Multivariate Bias Adjusted Tapered Predictive Process Models

Rajarshi Guhaniyogi

Department of Applied Mathematics & Statistics, University of California Santa Cruz, SOE 2, 1156 High Street, Santa Cruz, California 95064, USA

rguhaniy@ucsc.edu

Abstract

We extend earlier work on multivariate "low-rank" methods for the analysis of large multivariate spatial datasets. "Low-rank" methods usually operate on lower-dimensional subspaces and induce biases in the residual variance components as a result of over-smoothing or model mis-specification. Our current work attempts to characterize these biases, demonstrates their presence as a systemic phenomena, and explores remedial models without incurring computational costs. Our methodological contribution lies in the development of the multivariate tapered predictive process model that accounts for spatial correlations among multivariate components by the recently proposed *multivariate matern* correlation kernel. Both the proposed framework and the multivariate tapered predictive process model using linear model co-regionalization (LMC) ([1]) have been found to rectify bias in parameter estimation. We also prove novel theoretical results comparing *smoothness* properties of multivariate tapered predictive process models and classes of low rank models, including predictive processes. Finally, we illustrate our work using synthetic experiments as well as an application to forestry.

Keywords: Bayesian inference, covariance-tapering, low rank models, mean square differentiability, predictive process, spatial smoothness

1. Introduction

With the advent and expansion of Geographical Information Systems (GIS), along with related software, statisticians today routinely encounter large spatial or spatiotemporal datasets containing multiple variables observed across thousands of locations. This has, in turn, generated considerable interest in statistical modeling for location-referenced spatial data; see, for example, the books by [2, 3, 4] for a variety of methods and applications. Fitting Bayesian hierarchical spatial models to these datasets involves matrix decomposition of complexity n^3 with number of locations n for an univariate outcome. This makes computation of these models infeasible for large datasets. In popular spatial literature, this problem is generally referred to as the "big-n" problem. Evidently, multivariate and spatiotemporal processes exacerbate the problem.

Low rank or reduced rank models have become popular for analyzing large spatial and spatiotemporal datasets (see [5, 6, 7, 8]). The basic idea behind low rank models is to set a few "basis functions" in space, usually taken to be much smaller in number compared to the number of data locations, and to express the spatial process realizations over the entire set of observed locations in terms of only the few basis functions and associated coefficients. Depending on various choices of basis functions a wide class of low rank model have emerged in the recent past. In what follows, we will focus mostly on one specific class of low rank models, known as the predictive process model. In predictive process one considers a set of locations in the spatial domain, or "knots," and construct basis functions based on these knots. In this process, one must ensure the spatial information available from the entire set of locations can be summarized with the set of knots allowing some acceptable loss of information. This is achieved by pursuing a rich and flexible framework that integrates knot selection into the modeling ([9]).

Low rank spatial models have been widely deployed in the environmental and natural sciences to develop highly competent inferential frameworks for large spatial databases ([8, 7]). However, [10] report potential problems in pre-

diction and inference arising from the low-rank models. More specifically, [11] demonstrate how the *predictive process* yields biased estimates of certain variance components in spatial progeny trials and develops one remedy. Though multiple articles have been written on univariate spatial low-rank models, multivariate low rank models have found relatively less emphasis.

This manuscript embarks upon characterizing and understanding biases in the multivariate low rank models. Assuming the geostatistical model with Gaussian process prior on the spatial components as the "gold standard," we intend to investigate how low rank models approximate the "gold standard" in terms of spatial surface recovery and parameter estimation. In particular, it is observed that low rank models underestimate spatial variability and overestimate noise variability, thereby yielding "smoother" spatial surfaces. Such inaccuracies in terms of parameter estimation that results in smoothed surface estimation is referred to as *bias* in this article. We show their presence in low rank models as a systematic phenomenon, discuss their potential impact on spatial inference and explore remedies applicable to a wide range of hierarchical multivariate low-rank spatial process models. Our methodological contribution lies in proposing a multivariate extension of univariate tapered predictive process ([12]) that induces correlations among components by the multivariate matern kernel. We compare our proposed approach with the multivariate tapered predictive process model based on the linear model co-regionalization (LMC) ([1]) and found both of them as satisfactory tools for bias adjustment, with mixed relative performances. This article also highlights some of the desirable theoretical features of the multivariate tapered predictive process. In particular, we show that a class of low rank processes (including the predictive process) are infinitely mean square differentiable while tapered predictive process is mean square differentiable upto a certain order depending on the tapering correlation kernel and the parent Gaussian process. This is a novel theoretical result that concurs with the practical findings on the local behavior ("smoothness") of estimated spatial surfaces from these models. Such observations deem the choice of multivariate tapered predictive process attractive as a bias adjustment tool for multivariate

spatial applications.

The remainder of the article evolves as follows. Section 2 discusses low-rank spatial modeling in general, while Section 3 discusses how hierarchical Gaussian predictive process models help quantify bias in residual variability. Section 4 discusses multivariate tapered predictive process models to remedy such biases and also offers some theoretical comparisons. Section 5 talks about estimation and inference. Section 6 illustrates the different bias-adjusted models using two simulation studies, followed by a forestry application. Finally, Section 7 concludes the paper with some discussion and general conclusions.

2. Low-rank spatial models and related biases

2.1. Multivariate spatial process models-a brief review

Let $D \subset \Re^d$ be a subset of the d-dimensional Euclidean space and let $\boldsymbol{s} \in D$ be a generic point in D. In our subsequent application $d = 2$. Geostatistical multivariate settings typically envision, for each location $\boldsymbol{s} \in D$, an $m \times 1$ outcome or response $\boldsymbol{y}(\boldsymbol{s}) = (y_1(\boldsymbol{s}), y_2(\boldsymbol{s}), \ldots, y_m(\boldsymbol{s}))'$ whose mean is usually modeled as $\mathrm{E}[\boldsymbol{y}(\boldsymbol{s}) \mid \boldsymbol{X}(\boldsymbol{s}), \boldsymbol{\beta}, \boldsymbol{w}(\boldsymbol{s})] = \boldsymbol{X}(\boldsymbol{s})'\boldsymbol{\beta} + \boldsymbol{w}(\boldsymbol{s})$, where $\boldsymbol{X}(\boldsymbol{s})'$ is a $m \times p$ matrix of spatially referenced predictors (or completely known functions), whose i-th row is a $1 \times p$ vector of spatially referenced predictors $\boldsymbol{x}_i(\boldsymbol{s})'$, capturing large-scale variation or trends and $\boldsymbol{w}(\boldsymbol{s}) = (w_1(\boldsymbol{s}), w_2(\boldsymbol{s}), \ldots, w_m(\boldsymbol{s}))'$ is a $m \times 1$ vector of spatial process providing local adjustment (with structured dependence) to the mean, interpreted, often, as capturing the effect of unmeasured or unobserved covariates with a spatial pattern.

The customary process specification for $\boldsymbol{w}(\boldsymbol{s})$ is a zero-centered m-variate Gaussian process. Unless otherwise stated, in the rest of the article we will use uppercase bold and lowercase bold to denote matrices and vectors respectively. The process $\boldsymbol{w}(\boldsymbol{s})$ is completely specified by its *cross-covariance* function $\boldsymbol{C}_w(\boldsymbol{s}, \boldsymbol{t}; \boldsymbol{\Theta}_1)$, which, for any pair of locations $\boldsymbol{s}$ and $\boldsymbol{t}$, is an $m \times m$ matrix with $\mathrm{cov}\{w_i(\boldsymbol{s}), w_j(\boldsymbol{t})\}$ as its (i, j)-th element and $\boldsymbol{\Theta}_1$ is a collection of process parameters. This implies that for any finite set of n locations, say $\mathcal{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n\}$,

the $mn \times 1$ vector of realizations, $\boldsymbol{w} = (\boldsymbol{w}(\boldsymbol{s}_1)', \boldsymbol{w}(\boldsymbol{s}_2)', \ldots, \boldsymbol{w}(\boldsymbol{s}_n)')'$ follows a multivariate normal distribution with zero mean and a $mn \times mn$ blocked covariance matrix $\boldsymbol{C}_w(\boldsymbol{\Theta}_1)$ whose (i,j)-th block is given by the $m \times m$ matrix $\boldsymbol{C}_w(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\Theta}_1)$.

A valid multivariate process must ensure that $\boldsymbol{C}_w(\boldsymbol{\Theta}_1)$ is positive definite (hence symmetric too), which implies that $\boldsymbol{C}_w(\boldsymbol{s}, \boldsymbol{t}; \boldsymbol{\Theta}_1)$ must satisfy the following two conditions:

$$\begin{aligned} &\text{(i)} \quad \boldsymbol{C}_w(\boldsymbol{s}, \boldsymbol{t}; \boldsymbol{\Theta}_1) = \boldsymbol{C}_w(\boldsymbol{t}, \boldsymbol{s}; \boldsymbol{\Theta}_1)' \\ &\text{(ii)} \quad \textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{u}_i' \boldsymbol{C}_w(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\Theta}_1) \boldsymbol{u}_j > 0 \quad \forall \quad \boldsymbol{u}_i, \boldsymbol{u}_j \in \Re^m \setminus \{\boldsymbol{0}\} \, . \end{aligned} \tag{1}$$

The first condition in (1) ensures that $\boldsymbol{C}_w(\boldsymbol{\Theta}_1)$ is symmetric, although the cross-covariance matrix function itself need not be. The second condition in (1) ensures that $\boldsymbol{C}_w(\boldsymbol{\Theta}_1)$ is positive-definite. These must be satisfied for all integers n and any finite collection of locations $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\} \subset D$. Note that (1) implies that $\boldsymbol{C}_w(\boldsymbol{s}, \boldsymbol{s}; \boldsymbol{\Theta}_1)$ is precisely the variance-covariance matrix for the elements of $\boldsymbol{w}(\boldsymbol{s})$ within site $\boldsymbol{s}$. For a detailed review of how to construct valid cross-covariance matrix function, see [13, 14, 15, 16]. This article considers two different ways to construct cross-covariance matrix functions; linear model co-regionalization (LMC) and multivariate matern kernel.

2.1.1. Linear model co-regionalization

Assume, $v_i(\boldsymbol{s}) \sim GP(0, \rho(\boldsymbol{s}, \boldsymbol{t}; \boldsymbol{\theta}_i))$ and v_i's are independent over i. Following [17], the LMC approach assumes that $\boldsymbol{w}(\boldsymbol{s}) = \boldsymbol{A}\boldsymbol{v}(\boldsymbol{s})$, $\boldsymbol{v}(\boldsymbol{s}) = (v_1(\boldsymbol{s}), ..., v_m(\boldsymbol{s}))'$, which yields a highly structured cross-covariance function,

$$\boldsymbol{C}_w(\boldsymbol{s}, \boldsymbol{t}; \boldsymbol{\Theta}_1) = \boldsymbol{A}\boldsymbol{C}_v(\boldsymbol{s}, \boldsymbol{t}; \boldsymbol{\Theta}_1)\boldsymbol{A}' = \sum_{k=1}^{m} \boldsymbol{a}_k \boldsymbol{a}_k' \rho_k(\boldsymbol{s}_1, \boldsymbol{s}_2; \boldsymbol{\theta}_k) \, , \tag{2}$$

where $\boldsymbol{a}_k$ is the k-th column of $\boldsymbol{A}$, $\boldsymbol{\Theta}_1 = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_m\}$ (in univariate case $\boldsymbol{\Theta}_1 = \boldsymbol{\theta}_1$). A popular choice of $\rho(\cdot, \cdot)$ is the Matérn correlation function, with decay parameter θ_1 and smoothness parameter θ_2, given by

$$\rho(\boldsymbol{s}, \boldsymbol{t}; \boldsymbol{\theta}) = \frac{1}{2^{\theta_2 - 1}\Gamma(\theta_2)} (||\boldsymbol{s} - \boldsymbol{t}||\theta_1)^{\theta_2} \mathcal{K}_{\theta_2}(||\boldsymbol{s} - \boldsymbol{t}||; \theta_1); \;\; \theta_1 > 0, \; \theta_2 > 0 \tag{3}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)$. We denote the parent process under LMC as $w^{(1)}(\boldsymbol{s})$.

2.1.2. Multivariate matern kernel

Under LMC the smoothness of any component defaults to the roughest spatial process, thus the multivariate spatial process obtained by LMC does not admit components with different degrees of spatial smoothness, unless some structural zeroes are imposed on $\boldsymbol{A}$ ([14]). As a remedy we consider the multivariate matern correlation function. Though some variants of multivariate matern correlation function exist in the literature, we only consider the parsimonious multivariate matern function for the sake of simplicity. Here, the ith diagonal element of $\boldsymbol{C}_w(\boldsymbol{s}, \boldsymbol{t}, \boldsymbol{\Theta})$ is given by $C_{w,ii}(\boldsymbol{s}, \boldsymbol{t}, \boldsymbol{\Theta}) = \sigma_i^2 \rho(\boldsymbol{s}, \boldsymbol{t}, \theta_{1,i}, \theta_{2,i})$ and the (i, j)th off-diagonal entry is defined as $C_{w,ij}(\boldsymbol{s}, \boldsymbol{t}, \boldsymbol{\Theta}) = \zeta_{ij}\sigma_i\sigma_j\rho_1(\boldsymbol{s}, \boldsymbol{t}, \theta_{1,ij}, \theta_{2,ij})$, i.e. all the diagonal and off-diagonal entries are univariate matern covariance kernel with individual variance, scale and smoothness parameters. The parsimonious multivariate matern model ([18, 14]) assumes the following relationships between parameters

$$\begin{aligned}
&\theta_{2,ij} = (\theta_{2,i} + \theta_{2,j})/2, \ \forall \ i \neq j \\
&\theta_{1,1} = .. = \theta_{1,m} = \theta, \ \text{and} \ \theta_{1,ij} = \theta, \ \forall \ i \neq j \\
&\zeta_{ij} = \eta_{ij}(\theta_{2,i}\theta_{2,j})^{1/2}/[(\theta_{2,i} + \theta_{2,j})/2],
\end{aligned}$$

where $((\eta_{ij}))_{i,j=1}^m$ is a symmetric and non-negative definite matrix. The parent Gaussian process derived from the multivariate matern kernel is denoted by $w^{(2)}(\boldsymbol{s})$.

For the multivariate random noise, we assume $\boldsymbol{\epsilon}(\boldsymbol{s}_i) \overset{iid}{\sim} N(0, \boldsymbol{\Psi})$, $\boldsymbol{\Psi} = \text{diag}\ (\tau_1^2, ..., \tau_m^2)$ (when $m = 1$, we denote $\epsilon(\boldsymbol{s}_i) \overset{iid}{\sim} N(0, \tau^2)$). For an $mn \times 1$ vector of observed outcomes, $\boldsymbol{y} = (\boldsymbol{y}(\boldsymbol{s}_1), \boldsymbol{y}(\boldsymbol{s}_2), \ldots, \boldsymbol{y}(\boldsymbol{s}_n))'$ with a first stage conditionally independent Gaussian specification and associated priors, we construct a Bayesian hierarchical model

$$\begin{aligned}
p(\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{w} \,|\, \boldsymbol{y}) \propto p(\boldsymbol{\Theta}) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \times N(\boldsymbol{w} \,|\, \boldsymbol{0}, \boldsymbol{C}_w(\boldsymbol{\Theta}_1)) \\
\times \prod_{i=1}^{n} N(\boldsymbol{y}(\boldsymbol{s}_i) \,|\, \boldsymbol{X}(\boldsymbol{s}_i)'\boldsymbol{\beta} + \boldsymbol{w}(\boldsymbol{s}_i), \boldsymbol{\Psi}) \ , \qquad (4)
\end{aligned}$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_1, \boldsymbol{\Psi}\}$. The parameter $\boldsymbol{\Psi}$ (popularly known as *nugget* for the univariate case) captures multivariate random noise that may arise in the form of measurement error or micro-scale variability.

Spatial data analysis seeks to estimate the regression coefficients $\boldsymbol{\beta}$, which inform about large-scale trends and impact of predictors on the outcome, the unknown process parameters $\boldsymbol{\Theta}$, which inform about the nature of spatial associations and micro-scale variability, and the spatial effects $\boldsymbol{w}$ which inform about spatial patterns in the residual. Estimating (4) customarily proceeds using Markov chain monte carlo (MCMC) sampling. With Gaussian likelihoods, often we integrate out the spatial effects $\boldsymbol{w}$. This replaces the likelihood and the prior for $\boldsymbol{w}$ by $N(\boldsymbol{y} \mid \boldsymbol{X\beta}, \boldsymbol{C}_w + \boldsymbol{I}_n \otimes \boldsymbol{\Psi})$. In either case, estimation will involve $mn \times mn$ matrix decompositions, which will become exorbitant for large n.

2.2. Biases in low rank models

A computationally convenient alternative for modeling large multivariate spatial data is to replace the parent process in (4) with some linear combination of the form $\boldsymbol{w} \approx \boldsymbol{Z}_{1w}\boldsymbol{u}_1$, where $\boldsymbol{Z}_{1w}$ is an $mn \times mn^*$ matrix and $\boldsymbol{u}_1 = (u_1, u_2, \ldots, u_{mn^*})'$ is an $mn^* \times 1$ random vector. In semiparametric regression models using splines, $\boldsymbol{Z}_{1w}$'s are called the *basis functions* and u_j's the *basis coefficients*. Similarly, in the truncated Karhunen-Loeve expansion any row of $\boldsymbol{Z}_{1w}$ is the vector of truncated basis functions evaluated at a location and $\boldsymbol{u}_1$ is the corresponding vector of coefficients. It is be noted the Parent Gaussian process (4) when represented under the same basis function gives rise to $\boldsymbol{w} = [\boldsymbol{Z}_{1w} : \boldsymbol{Z}_{2w}]\boldsymbol{u}$, where $\boldsymbol{Z}_{2w}$ is an $mn \times m(n-n^*)$ matrix and $\boldsymbol{u} = (\boldsymbol{u}_1', \boldsymbol{u}_2')'$.

Irrespective of their precise specifications, low-rank models tend to overestimate the residual variance. This bias arises from systemic over-smoothing or model under-specification by the low-rank model when compared to the parent model. In fact, this becomes especially transparent from writing the parent likelihood and the low-rank likelihood as mixed linear models (see [19]),

$$\textbf{Parent likelihood:} \quad \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_{1w}\boldsymbol{u}_1 + \boldsymbol{Z}_{2w}\boldsymbol{u}_2 + \boldsymbol{\epsilon}_1\,;\ \boldsymbol{\epsilon}_1 \sim N(\boldsymbol{0}, \boldsymbol{I} \otimes \boldsymbol{\Psi});$$

$$\textbf{Low rank likelihood:} \quad \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_{1w}\boldsymbol{u}_1 + \boldsymbol{\epsilon}_2\,; \quad \boldsymbol{\epsilon}_2 \sim N(\boldsymbol{0}, \boldsymbol{I} \otimes \boldsymbol{\Psi})\ .$$

For fixed $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}_1$, the basis functions forming the columns of $\boldsymbol{Z}_{2w}$ in the parent likelihood are absorbed into the residual error in the low rank likelihood, leading to an upward bias in the estimate of the nugget. More precisely, letting $\boldsymbol{P}_{\boldsymbol{Z}_w} = \boldsymbol{Z}_w(\boldsymbol{Z}'_w\boldsymbol{Z}_w)^{-1}\boldsymbol{Z}'_w$ (the orthogonal projection matrix or "hat" matrix into the column space of $\boldsymbol{Z}_w$), standard linear model calculations reveal the residual variability from the parent model is quantified by $(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}_w})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$, while that from the low-rank model is given by $(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}_{1w}})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$. Using the fact that $\boldsymbol{P}_{\boldsymbol{Z}_w} = \boldsymbol{P}_{\boldsymbol{Z}_{1w}} + \boldsymbol{P}_{[(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}_{1w}})\boldsymbol{Z}_{2w}]}$, the excess residual variability in the low-rank likelihood appears as $(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{P}_{[(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}_{1w}})\boldsymbol{Z}_{2w}]}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$.

Although this excess residual variability can be quantified as above, it is less clear how the low-rank spatial likelihood could be modified to compensate for this overestimation without adding significantly to the computational burden. To characterize this bias and provide a computationally feasible remedy for hierarchical models, it will be helpful to work with a low-rank spatial process rather than a low-rank likelihood. In this context, we remark that not all low-rank models lead to a straightforward quantification of this excess residual variability (or bias). For instance, low-rank models based upon kernel convolutions ([5]) approximate $\boldsymbol{w}(\boldsymbol{s})$ with $\boldsymbol{w}_{KC}(\boldsymbol{s}) = \sum_{j=1}^{n^*}\boldsymbol{k}(\boldsymbol{s}-\boldsymbol{s}^*_j)u_j$, where $\boldsymbol{k}(\cdot)$ is a $m \times 1$ vector of kernel functions and $u_j \overset{iid}{\sim} N(0,1)$, assumed to arise from a Brownian motion $U(\boldsymbol{v})$ on $\Re^2$. This yields

$$\boldsymbol{w}(\boldsymbol{s})-\boldsymbol{w}_{KC}(\boldsymbol{s}) = \int \boldsymbol{k}(\boldsymbol{s}-\boldsymbol{v})dU(\boldsymbol{v})-\sum_{j=1}^{n^*}\boldsymbol{k}(\boldsymbol{s}-\boldsymbol{s}^*_j)u_j \approx \sum_{j=n^*+1}^{\infty}\boldsymbol{k}(\boldsymbol{s}-\boldsymbol{s}^*_j)u_j\ , \quad (5)$$

which does not, in general, render a closed form and may be difficult to compute accurately. Expression for the residual process is, however, simplified for a special class of low-rank models, known as, predictive process models. We discuss predictive process models in the next section.

3. Multivariate predictive process models

An optimal projection of the process $\boldsymbol{w}(\boldsymbol{s})$ at a generic location $\boldsymbol{s}$, based upon its realization over $\mathcal{S}^*$, is given by the "kriging equation" $\boldsymbol{w}_{pp}(\boldsymbol{s}) = \mathrm{E}\left[\boldsymbol{w}(\boldsymbol{s}) \,|\, \boldsymbol{w}^*\right]$, where $\boldsymbol{w}^* = (\boldsymbol{w}(\boldsymbol{s}_1^*)', \boldsymbol{w}(\boldsymbol{s}_2^*)', \ldots, \boldsymbol{w}(\boldsymbol{s}_{n^*}^*)')'$. [8] refer to $\boldsymbol{w}_{pp}(\boldsymbol{s})$ as the *predictive process* derived from the *parent process* $\boldsymbol{w}(\boldsymbol{s})$. When the parent process is a zero-centered Gaussian process with covariance function $\boldsymbol{C}_w(\boldsymbol{s}_1, \boldsymbol{s}_2)$, we can write the predictive process as $\boldsymbol{w}_{pp}(\boldsymbol{s}) = \mathrm{E}\left[\boldsymbol{w}(\boldsymbol{s}) \,|\, \boldsymbol{w}^*\right] = \mathcal{C}_w(\boldsymbol{s}, \mathcal{S}^*)' \boldsymbol{C}_w^{*-1} \boldsymbol{w}^*$, where $\mathcal{C}_w(\boldsymbol{s}, \mathcal{S}^*)'$ is the $m \times mn^*$ block matrix composed of $m \times m$ block matrix $\boldsymbol{C}_w(\boldsymbol{s}, \boldsymbol{s}_j^*)$, $j = 1, .., n^*$ and $\boldsymbol{C}_w^*$ is the $mn^* \times mn^*$ covariance matrix with elements $\boldsymbol{C}_w(\boldsymbol{s}_i^*, \boldsymbol{s}_j^*)$. Since $\boldsymbol{w}^*$ is multivariate normal with zero mean and $mn^* \times mn^*$ variance-covariance matrix $\boldsymbol{C}_w^*$, the predictive process is itself a nonstationary Gaussian process arising from a spatially adaptive linear transformation of the parent process over the set of knots. Replacing $\boldsymbol{w}(\boldsymbol{s})$ with $\boldsymbol{w}_{pp}(\boldsymbol{s})$ in (4), leads to its predictive process counterpart

$$p(\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{w}^* \,|\, \boldsymbol{y}) \propto p(\boldsymbol{\Theta}) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \times N(\boldsymbol{w}^* \,|\, \boldsymbol{0}, \boldsymbol{C}_w^*) \times \prod_{i=1}^{n} N(\boldsymbol{y}(\boldsymbol{s}_i) \,|\, \boldsymbol{X}(\boldsymbol{s}_i)'\boldsymbol{\beta} + \boldsymbol{w}_{pp}(\boldsymbol{s}_i), \boldsymbol{\Psi}). \tag{6}$$

Computational gains are achieved since matrix computations now involve the $mn^* \times mn^*$ matrix $\boldsymbol{C}_w^*$, where n^* is chosen to be much smaller than n.

The predictive process is a low-rank process and its partial realizations produce a low-rank likelihood. Being smoother than the parent process, it tends to have lower variance which, in turn, inflates the residual variability often manifested as an overestimation of $\boldsymbol{\Psi}$. In fact, for fixed $\mathcal{S}^*$ we have, $\boldsymbol{w}(\cdot) - \boldsymbol{w}_{pp}(\cdot) \sim GP(\boldsymbol{0}, \boldsymbol{\Delta}(\cdot))$, where,

$$\boldsymbol{\Delta}(\boldsymbol{s}, \boldsymbol{t}) = \boldsymbol{C}_w(\boldsymbol{s}, \boldsymbol{t}) - \mathcal{C}_w(\boldsymbol{s}, \mathcal{S}^*)' \boldsymbol{C}_w^{*-1} \mathcal{C}_w(\boldsymbol{t}, \mathcal{S}^*)\ , \tag{7}$$

and $\boldsymbol{\Delta}(\boldsymbol{s}, \boldsymbol{s})$ being abbreviated as $\boldsymbol{\Delta}(\boldsymbol{s})$. Such closed form expression facilitates deriving a bound of the stochastic error incurred due to the approximation of the Gaussian process through a predictive process.

A remedy for the bias in the predictive process model ([11, 20]) is to use $\boldsymbol{w}_{mpp}(\boldsymbol{s}) = \boldsymbol{w}_{pp}(\boldsymbol{s}) + \boldsymbol{\epsilon}_{mpp}(\boldsymbol{s})$, often called the *modified predictive process*, where $\boldsymbol{\epsilon}_{mpp}(\boldsymbol{s}) \overset{ind}{\sim} N(\boldsymbol{0}, \boldsymbol{\Delta}(\boldsymbol{s}))$ and $\boldsymbol{\epsilon}_{mpp}(\boldsymbol{s})$ is independent of $\boldsymbol{w}_{pp}(\boldsymbol{s})$. Now, the variance of $\boldsymbol{w}_{mpp}(\boldsymbol{s})$ equals that of the parent process $\boldsymbol{w}(\boldsymbol{s})$.

The marginalized modified predictive process is estimated by

$$p(\boldsymbol{\beta}, \boldsymbol{\Theta} \,|\, \boldsymbol{y}) \propto p(\boldsymbol{\Theta}) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \times N(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{y};mpp}) \;, \tag{8}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{y};mpp} = \mathcal{C}_w' \boldsymbol{C}_w^{*-1} \mathcal{C}_w + \boldsymbol{D}_{mpp}$, $\mathcal{C}_w'$ is the $mn \times mn^*$ matrix whose (i,j)-th element is $\mathcal{C}_w(\boldsymbol{s}_i, \boldsymbol{s}_j^*)$ and $\boldsymbol{D}_{mpp}$ is a block diagonal matrix with (i,i)-th block diagonal element $\boldsymbol{\Delta}(\boldsymbol{s}_i) + \boldsymbol{\Psi}$. Estimation involves the inverse and determinant of $\boldsymbol{\Sigma}_{\boldsymbol{y};mpp}$, which can be efficiently computed using the Sherman-Woodbury-Morrison formula (e.g. [20]). Next section will undertake an exploration of more sophisticated multivariate model based bias adjustment. One important point to make before we conclude this section is that both predictive process and modified predictive process are dependent on the choice of the multivariate covariance kernel. Hereon, we refer to $\boldsymbol{w}_{pp}^{(1)}$ and $\boldsymbol{w}_{pp}^{(2)}$ as the predictive process obtained from LMC and multivariate matern kernel respectively. $\boldsymbol{w}_{mpp}^{(1)}$ and $\boldsymbol{w}_{mpp}^{(2)}$ are defined analogously.

4. Multivariate tapered adjustment to predictive process models

As an alternative to the modified predictive process, a modification accruing computational benefits, *tapers* the covariance kernel of the process $\boldsymbol{w}(\boldsymbol{s}) - \boldsymbol{w}_{pp}(\boldsymbol{s})$. Tapering covariance kernels offer an alternative means of dimension reduction, by producing sparse spatial covariance matrices and have received much attention in the recent past ([21, 22, 12]). To make our presentation more clear, we initially focus on univariate ($m = 1$) tapering. Later we will generalize it in multivariate processes.

The underlying idea of tapering is to use a compactly supported covariance function ([23]) as a *tapering kernel* $C_\nu(\boldsymbol{s}_1, \boldsymbol{s}_2)$, which is a positive-definite function satisfying

$$C_\nu(\boldsymbol{s}_1, \boldsymbol{s}_2) = 0 \quad \text{if} \quad ||\boldsymbol{s}_1 - \boldsymbol{s}_2|| > \nu \;,$$

where ν is the distance beyond which the covariance becomes zero. Now consider a covariance function obtained by taking product of $C_\nu(\cdot,\cdot)$ and any spatial covariance function $C_w(\cdot,\cdot)$. Then $C_{tap}(\cdot,\cdot) = C_\nu(\cdot,\cdot)C_w(\cdot,\cdot)$. Tapering proceeds to seek approximate inference by replacing covariance matrices based on $C_w(\cdot,\cdot)$ by those based on $C_{tap}(\cdot,\cdot)$. Such an approximation is justified and provides asymptotically efficient results as suggested by [24] and [25].

From an implementation standpoint, tapering introduces a sparse structure for the dispersion matrix from the Gaussian process model. Referring to the univariate ($m = 1$) case, let $\boldsymbol{T}$ denote the $n \times n$ matrix with (i,j)-th element $C_\nu(\boldsymbol{s}_i, \boldsymbol{s}_j)$. Clearly the matrix $\boldsymbol{T}$ will have zero entries for any pair of locations separated by more than ν units and, hence, is sparse. There are choices aplenty for the tapering kernel, but the more widely used kernels use the Wendland family of tapered covariance functions ([21]). One particularly popular choice is given by $C_\nu(\boldsymbol{s}_1, \boldsymbol{s}_2) = \left(1 - \frac{h}{\nu}\right)_+^4 \left(1 + 4\frac{h}{\nu}\right)$, where $h = ||\boldsymbol{s}_1 - \boldsymbol{s}_2||$. Note that ν is typically not estimated, but fixed to achieve the desired degree of sparsity in the dispersion matrix ([22]). In a Bayesian context, ν can possibly be estimated using some prior distribution, but such priors will need to be strongly informative. We avoid this needless complexity and work with a fixed ν in the subsequent development.

Tapered covariance structures have been used effectively for analyzing large spatial datasets. In particular, [12, 1] propose tapering the residual process for accurate small scale and large scale spatial variations. This article explores two multivariate version of the tapered predictive process with an eye to reduce bias in the estimation of error variances. **Strategy 1** finds mention in [1], while **Strategy 2** is a new method proposed in this article.

Strategy 1: we use the popularly known LMC described in Section 2.1.1 to model the correlations between components of the multivariate spatial process as in $\boldsymbol{w}^{(1)}(\boldsymbol{s}) = \boldsymbol{A}\boldsymbol{v}(\boldsymbol{s})$. We then use the predictive process $\boldsymbol{v}_{pp}(\boldsymbol{s})$ of $\boldsymbol{v}(\boldsymbol{s})$ as in (6), and taper the covariance kernel of the residual process $\boldsymbol{v}(\boldsymbol{s}) - \boldsymbol{v}_{pp}(\boldsymbol{s})$. More precisely, assume $\boldsymbol{\eta}(\boldsymbol{s}) = (\eta_1(\boldsymbol{s}), ..., \eta_m(\boldsymbol{s}))'$ is a multivariate spatial random

field independent of $\boldsymbol{w}(\boldsymbol{s})$ with $\eta_k(\boldsymbol{s}) \sim \text{GP}(0, C_{\nu_k}(\cdot,\cdot))$ independently over $k = 1, ..., m$. Evidently, $\boldsymbol{v}(\boldsymbol{s}) - \boldsymbol{v}_{pp}(\boldsymbol{s})$ follows a Gaussian process with a covariance kernel $\boldsymbol{C}_{res,v}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \boldsymbol{C}_v(\boldsymbol{s}_1, \boldsymbol{s}_2) - \mathcal{C}_v(\boldsymbol{s}_1, \mathcal{S}^*)'\boldsymbol{C}_v^{*-1}\mathcal{C}_v(\boldsymbol{s}_2, \mathcal{S}^*)$. The tapered predictive process is defined to be,

$$\boldsymbol{w}_{tap}^{(1)}(\boldsymbol{s}) = \boldsymbol{w}_{pp}^{(1)}(\boldsymbol{s}) + \boldsymbol{A}[(\boldsymbol{v}(\boldsymbol{s}) - \tilde{\boldsymbol{v}}(\boldsymbol{s})) \odot \boldsymbol{\eta}(\boldsymbol{s})] \tag{9}$$

which can easily be seen as a Gaussian process with covariance function $\boldsymbol{C}_{tap,w}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \boldsymbol{A}\boldsymbol{C}_{res,v}(\boldsymbol{s}_1, \boldsymbol{s}_2)\boldsymbol{C}_\nu(\boldsymbol{s}_1, \boldsymbol{s}_2)\boldsymbol{A}' + \boldsymbol{C}_{w_{pp}}(\boldsymbol{s}_1, \boldsymbol{s}_2)$.

These specifications yield $N(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{y};tap,1})$ as the likelihood in (8), where

$$\begin{aligned} \boldsymbol{D}_{tap,1} &= \mathcal{A}\left[\left(\boldsymbol{C}_v - \mathcal{C}_v'\boldsymbol{C}_v^{*-1}\mathcal{C}_v\right) \odot \boldsymbol{T}\right]\mathcal{A}' \\ \Sigma_{\boldsymbol{y},tap,1} &= \mathcal{C}_w'\boldsymbol{C}_w^{*-1}\mathcal{C}_w + \boldsymbol{D}_{tap,1} + \boldsymbol{I}_n \otimes \boldsymbol{\Psi}\ , \end{aligned} \tag{10}$$

$\mathcal{A} = \boldsymbol{I} \otimes \boldsymbol{A}$, $\boldsymbol{T} = \{\text{diag}(C_{\nu_k}(\boldsymbol{s}_1, \boldsymbol{s}_j))_{k=1}^m\}_{i,j=1}^n$. This dispersion matrix is $mn \times mn$, but the sparse structure can be utilized, again in conjunction with the Sherman-Woodbury-Morrison formula, to achieve substantial computational gains.

Strategy 2: We use multivariate matern kernel described in Section 2.1.2 to model componentwise correlations in the parent process $\boldsymbol{w}^{(2)}(\boldsymbol{s})$. Then after we taper the correlation kernel of the residual process $\boldsymbol{w}^{(2)}(\boldsymbol{s}) - \boldsymbol{w}_{pp}^{(2)}(\boldsymbol{s})$. More precisely, with the abuse of notation, assume $\boldsymbol{\eta}(\boldsymbol{s}) = (\eta_1(\boldsymbol{s}), ..., \eta_m(\boldsymbol{s}))' \sim GP(\boldsymbol{0}, \rho_{tap}(\cdot,\cdot))$, where $\rho_{tap}(\boldsymbol{s}_i, \boldsymbol{s}_j) = [r_{k_1,k_2}C_{\nu_{k_1,k_2}}(\boldsymbol{s}_i, \boldsymbol{s}_j)]_{k_1,k_2=1}^m$, with $r_{k_1,k_2}, \nu_{k_1,k_2}$, $k_1, k_2 = 1, ..., m$ being a set of numbers such that $\nu_{k_1,k_2} > 0$. Clearly, the multi-taper function tapers both covariance and cross-covariance. [26] provides a number of ways to choose $r_{k_1,k_2}, \nu_{k_1,k_2}$ and C_ν, $k_1, k_2 = 1, ..., m$ to arrive at valid multivariate tapering kernels. In this article we choose the simple separable multi-taper discussed in [26]. To elaborate it further, choose $r_{k_1,k_1} = 1$ and $-1 \leq r_{k_1,k_2} \leq 1$, $k_1, k_2 = 1, .., m$ so that $((r_{k_1,k_2}))_{k_1,k_2=1}^m$ becomes a symmetric positive semi-definite matrix. C_ν is employed as in Strategy 1, $C_\nu(\boldsymbol{s}_1, \boldsymbol{s}_2) = \left(1 - \frac{h}{\nu}\right)_+^4 \left(1 + 4\frac{h}{\nu}\right)$. $\nu_{k_1,k_2} = \nu$, $\forall k_1, k_2$, determines the radius of the ball in $\mathbb{R}^2$ over which the correlation kernel is compactly supported. Simulation studies (Section 6) discusses various possible choices of r_{k_1,k_2} and their

impact on performance. The tapered predictive process under the multivariate matern kernel is given by

$$\boldsymbol{w}_{tap}^{(2)} = \boldsymbol{w}_{pp}^{(2)}(\boldsymbol{s}) + (\boldsymbol{w}^{(2)}(\boldsymbol{s}) - \boldsymbol{w}_{pp}^{(2)}(\boldsymbol{s}))\boldsymbol{\eta}(\boldsymbol{s}) \tag{11}$$

Define, $\boldsymbol{J}_{ij}$ as a the (i,j)th block of dimension $m \times m$ for an $mn \times mn$ matrix $\boldsymbol{J}$. Let the (k_1, k_2)th entry of $\boldsymbol{J}_{ij}$ be given by $r_{k_1,k_2} C_{\nu_{k_1,k_2}}(\boldsymbol{s}_i, \boldsymbol{s}_j)$. This specification yields a likelihood of $N(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{y};tap,2})$, where

$$\begin{aligned} \boldsymbol{D}_{tap,2} &= \left(\boldsymbol{C}_w - \mathcal{C}_w' \boldsymbol{C}_w^{*-1} \mathcal{C}_w\right) \odot \boldsymbol{J} \\ \Sigma_{\boldsymbol{y},tap,2} &= \mathcal{C}_w' \boldsymbol{C}_w^{*-1} \mathcal{C}_w + \boldsymbol{D}_{tap,2} + \boldsymbol{I}_n \otimes \boldsymbol{\Psi}. \end{aligned} \tag{12}$$

Although [27] proposes tapering the parent Gaussian process under the multivariate matern kernel, the **Strategy 2** proposed by us is the first to suggest tapering the residual process under multivariate matern kernel.

We point out that the modified predictive process did not account for the cross-covariance in the residual process $\boldsymbol{w}(\boldsymbol{s}) - \boldsymbol{w}_{pp}(\boldsymbol{s})$. Instead, it only adjusted for the variability in this residual using the independent process $\boldsymbol{\epsilon}_{mpp}(\boldsymbol{s})$. The tapered approach, on the other hand, accounts for the residual spatial association as well as the variability. Putting $\nu_k = 0$ and $\nu_k = \infty \; \forall \, k$ in the tapered predictive process in **Strategy 1** yields the modified predictive process and the parent Gaussian process respectively. Similarly, $\nu_{k_1,k_2} = 0$ and $\nu_{k_1,k_2} = \infty$ $\forall \, k_1, k_2$ leads to the modified predictive process and parent Gaussian process in **Strategy 2**. It is also quite evident from (9) that while the predictive process approximation of $\boldsymbol{w}(\boldsymbol{s})$ only offers a good approximation of the original covariance function at long distances due to the disappearance of the residual term, $\boldsymbol{w}_{tap}(\boldsymbol{s})$ aims at approximating the original covariance function both at long and short distances. In this sense, tapering enriches the approximation to the underlying parent process.

4.1. Dispersion matrix distances

Given the primary aim of the above three model-based strategies for modeling low-rank spatial covariance matrices is to provide good approximation of

the parent Gaussian process, it may be instructive to ascertain each of their "distances" from the parent model's dispersion matrix. Let $\mathbf{\Sigma}_{\boldsymbol{y};pp,1}$, $\mathbf{\Sigma}_{\boldsymbol{y};mpp,1}$ and $\mathbf{\Sigma}_{\boldsymbol{y};tap,1}$ be the marginal dispersion matrices corresponding to the predictive process, the modified predictive process and the tapered adjustment respectively for LMC with strategy 1 and $\mathbf{\Sigma}_{\boldsymbol{y};pp,2}$, $\mathbf{\Sigma}_{\boldsymbol{y};mpp,2}$ and $\mathbf{\Sigma}_{\boldsymbol{y};tap,2}$ be the marginal dispersion matrices of the same using strategy 2 respectively. Further assume $\mathbf{\Sigma}_{\boldsymbol{y};full,1}$ and $\mathbf{\Sigma}_{\boldsymbol{y};full,2}$ are the marginal covariance matrices of the parent Gaussian process with LMC and multivariate matern respectively. We use the following metric,

$$\Delta_{*}^{(i)} = \|\Sigma_{\boldsymbol{y};*,i} - \Sigma_{\boldsymbol{y};full,i}\|_2 \ , i = 1, 2,$$

where $||*||_2$ is the l_2 matrix norm, to compute $\Delta_{tap}^{(i)}$, $\Delta_{mpp}^{(i)}$ and $\Delta_{pp}^{(i)}$. We now have the following lemma.

Lemma 4.1. *Let $\Delta_{pp}^{(1)}$, $\Delta_{mpp}^{(1)}$ and $\Delta_{tap}^{(1)}$ be the three metrics defined above. Then, for any fixed $\mathbf{\Theta}$ we have the following inequality:*

$$\Delta_{pp}^{(1)} \geq \Delta_{mpp}^{(1)} \geq \Delta_{tap}^{(1)} \, .$$

Proof See Appendix A.

Remark 1: Lemma 4.1 holds even when different tapering kernels are used for different marginal covariance functions for different components.

Lemma 4.2. *Let $\Delta_{pp}^{(2)}$, $\Delta_{mpp}^{(2)}$ and $\Delta_{tap}^{(2)}$ be the three metrics defined above. Then, for any fixed $\mathbf{\Theta}$ we have the following inequality:*

$$\Delta_{pp}^{(2)} \geq \Delta_{mpp}^{(2)} \geq \Delta_{tap}^{(2)} \, .$$

The proof follows exactly along the line of the arguments provided for the proof of Lemma 4.1. The results provide one justification to use the multivariate tapered predictive process as a better approximation to the parent Gaussian process model compared to the modified predictive process.

4.2. Spatial Smoothness

This section devotes to study smoothness properties of general low rank models along with the tapered predictive process model. All results are presented for univariate spatial processes whereupon smoothness properties for multivariate processes follow from componentwise smoothness. For the sake of definiteness, we assume $v(\boldsymbol{s}) \sim GP(0, \rho(\cdot,\cdot,\theta_1,\theta_2))$ and $w(\boldsymbol{s}) = \sigma v(\boldsymbol{s})$. The entire section studies smoothness properties of various classes of models through mean square properties of the same.

Mean square properties of a random field provide a way to formalize the concept of spatial smoothness by introducing the idea of derivatives of random fields. Derivatives of the random fields have been discussed in [28], [29] and [30]. Let $w(\boldsymbol{s})$ be a real valued spatial process, the process $\{w(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{R}^2\}$ is L_2 continuous at $\boldsymbol{s}_0$ if $\lim_{\boldsymbol{s}\to\boldsymbol{s}_0} E\{w(\boldsymbol{s}) - w(\boldsymbol{s}_0)\}^2 = 0$. Moreover, if $w(\boldsymbol{s})$ is stationary with $\text{cov}\{w(\boldsymbol{s}), w(\boldsymbol{s}')\} = C_w(\boldsymbol{s} - \boldsymbol{s}')$ then the process $w(\boldsymbol{s})$ is mean square continuous at all sites $\boldsymbol{s}$ if C_w is continuous at $\mathbf{0}$. Analogous to the definition of mean square continuity, mean square differentiability of a process $w(\boldsymbol{s})$ demands the existence of a vector $\nabla w(\boldsymbol{s}_0)$, known as the gradient vector, such that, for any unit vector $\boldsymbol{u}$,

$$\lim_{h\to 0} \text{E}\left\{\frac{w(\boldsymbol{s}_0 + h\boldsymbol{u}) - w(\boldsymbol{s}_0)}{h} - \langle \nabla w(\boldsymbol{s}_0), \boldsymbol{u}\rangle\right\}^2 = 0 \tag{13}$$

Let $\nabla w(\boldsymbol{s}_0) = (\nabla w_1(\boldsymbol{s}_0), \nabla w_2(\boldsymbol{s}_0))'$; the process $w(\cdot)$ is said to be twice mean square differentiable at $\boldsymbol{s}_0$ if each of $\nabla w_i(\boldsymbol{s}_0)$ is mean square differentiable at $\boldsymbol{s}_0$. Let the mean square derivatives of $\nabla w_1(\boldsymbol{s}_0)$ and $\nabla w_2(\boldsymbol{s}_0)$ be $\nabla^2 w_1(\boldsymbol{s}_0) = (\nabla^2 w_{11}(\boldsymbol{s}_0), \nabla^2 w_{12}(\boldsymbol{s}_0))'$ and $\nabla^2 w_2(\boldsymbol{s}_0) = (\nabla^2 w_{21}(\boldsymbol{s}_0), \nabla^2 w_{22}(\boldsymbol{s}_0))$ respectively; stacking them together the second mean square derivative of $w(\boldsymbol{s}_0)$ is given by $\nabla^2 w(\boldsymbol{s}_0) = ((\nabla^2 w_1(\boldsymbol{s}_0))', (\nabla^2 w_2(\boldsymbol{s}_0))')'$. Continuing in this way, we denote the m_1-th order mean square derivative by $\nabla^{m_1} w(\boldsymbol{s}_0)$, obtained by stacking all the 2^{m_1} elements of the set $\mathcal{P}_{m_1} = \left\{\nabla^{m_1} w_{i_1,\ldots,i_{m_1}}(\boldsymbol{s}_0) : (i_1,\ldots,i_{m_1}) \in \{1,2\}^{m_1}\right\}$.

With this definition, when $w(\boldsymbol{s})$ is stationary, mean square differentiability only requires the existence of the 2nd order derivative of $C_w(\cdot)$ at 0. On a similar

note, existence of $C_w^{(2m_1)}(0)$ ensures m_1 times mean square differentiability of a stationary process $w(\boldsymbol{s})$.

In the light of the aforesaid concepts it is always instructive to compare general low rank models (predictive process being a special case) and tapered predictive process models. We first prove the result comparing smoothness of predictive process, modified predictive process and tapered predictive process.

Theorem 4.3. *Let, $C_w(\boldsymbol{s}, \boldsymbol{t})$ be a Matern correlation function with $m_1 < \theta_2 < m_1 + 1$ and assume the tapering kernel $C_\nu(\cdot)$ is $2k$ times differentiable at $\mathbf{0}$. Then,*

1. *The predictive process model is infinitely mean square differentiable except at the set of knot points $\mathcal{S}^*$.*
2. *The modified predictive process is not even mean square continuous at any point.*
3. *The tapered predictive process is $\min(m_1, k)$-times mean square differentiable except at the set of knot points $\mathcal{S}^*$.*

Proof See Appendix A.

Remark 3: Following the proof of Theorem 4.3, it can be argued that any low rank model that replaces the parent Gaussian process $w(\boldsymbol{s})$ with a low rank model $\tilde{w}(\boldsymbol{s}) = \boldsymbol{K}(\boldsymbol{s})\boldsymbol{u}$ that uses basis functions $\boldsymbol{K}(\boldsymbol{s})$ with basis coefficients $\boldsymbol{u}$, is infinitely differentiable provided the function $\boldsymbol{K}(\boldsymbol{s})$ is smooth (infinitely differentiable).

Theorem 4.3 indicates the continuity and differentiability of the parent process $w(\cdot)$ can be retained in the tapered predictive process if the tapering kernel can be suitably chosen. Therefore, the tapered predictive approach facilitates spatial gradient assessment of the spatial surface for large datasets quite accurately as it can ensure a similar degree of smoothness to the parent Gaussian process while being computationally more efficient. As an example, in the study of spatial gradients, [31] employs Matérn correlation kernel with $\theta_2 = \frac{3}{2}$ which yields once differentiable spatial realizations. It is discussed in [23] that the

Wendland tapering kernel $C_\nu(\boldsymbol{s},\boldsymbol{t}) = (1-\frac{\|\boldsymbol{s}-\boldsymbol{t}\|}{\nu})_+^4(1+4\frac{\|\boldsymbol{s}-\boldsymbol{t}\|}{\nu})$ is twice differentiable at $\mathbf{0}$ which implies that $\eta(\boldsymbol{s})$ is once continuously differentiable. Referring to the Theorem 4.3, it is now quite straightforward to see that the tapered predictive process yields once differentiable spatial realization. In this sense, tapered predictive process, even with a computationally efficient structure, offers important assessment of spatial gradients. In a future article, we propose to investigate spatial gradients with tapered predictive process models in details.

Before concluding this section a few remarks on multivariate tapered predictive processes are in order. Note that multivariate tapered predictive process with strategy 1 (LMC) is derived from the parent multivariate Gaussian process $\boldsymbol{w}(\boldsymbol{s})$ with its ith component given by $w_i(\boldsymbol{s}) = \sum_{k=1}^m a_{ik}v_k(\boldsymbol{s})$. Proposition 1 in the Appendix can be invoked to argue that LMC leads to $w_i(\boldsymbol{s})$'s having the same degree of smoothness, even if $v_k(\boldsymbol{s})$'s have differential smoothness. Things are improved by introducing structural zeroes to $\boldsymbol{A}$. Mutivariate matern model (strategy 2), on the other hand, is equipped to model various degrees of smoothness in individual components. Therefore, theoretically greater flexibility is achieved through multivariate matern model.

5. Inference & Model fit

Once the parameters have been estimated, inferential interest turns to spatial prediction. Here, a few situations are of interest. Let $\boldsymbol{s}_0$ be any location in the domain, where we seek to predict $\boldsymbol{y}(\boldsymbol{s}_0)$, based on a given matrix of predictors $\boldsymbol{X}(\boldsymbol{s}_0)'$. For the marginalized model, spatial prediction proceeds from the posterior predictive distribution

$$p(\boldsymbol{y}(\boldsymbol{s}_0)\,|\,\boldsymbol{y}) = \int p(\boldsymbol{y}(\boldsymbol{s}_0)\,|\,\boldsymbol{y},\boldsymbol{\Theta})p(\boldsymbol{\Theta}\,|\,\boldsymbol{y})\,d\boldsymbol{\Theta}. \tag{14}$$

Posterior predictive sampling is achieved using *composition*. For each $\{\boldsymbol{\Theta}^{(l)}\}$, $l = 1,2,\ldots,L$, obtained from the posterior distribution $p(\boldsymbol{\Theta}\,|\,\boldsymbol{y})$, we draw $\boldsymbol{y}(\boldsymbol{s}_0)^{(l)}$ from $p(\boldsymbol{y}(\boldsymbol{s}_0)\,|\,\boldsymbol{y},\boldsymbol{\Theta}^{(l)})$. The resulting $\boldsymbol{y}(\boldsymbol{s}_0)^{(l)}$, $l = 1,2,\ldots,L$ are samples from (14). This is especially simple for Gaussian likelihoods because $p(\boldsymbol{y}(\boldsymbol{s}_0)\,|\,\boldsymbol{y},\boldsymbol{\Theta})$ then turns out to be normal distribution.

Bayesian inference is especially attractive for spatial data analysis because it facilitates full inference on the latent spatial processes $\boldsymbol{w}_*(\boldsymbol{s})$ or even $\boldsymbol{w}(\boldsymbol{s})$. For example, the posterior distribution for $\boldsymbol{w}_*(\boldsymbol{s}_0)$ for an arbitrary location $\boldsymbol{s}_0$ is given by,

$$p(\boldsymbol{w}_*(\boldsymbol{s}_0) \,|\, \boldsymbol{y}) = \int p(\boldsymbol{w}_*(\boldsymbol{s}_0) \,|\, \boldsymbol{w}^*, \boldsymbol{\Theta}) p(\boldsymbol{w}^* \,|\, \boldsymbol{y}, \boldsymbol{\Theta}) p(\boldsymbol{\Theta} \,|\, \boldsymbol{y})\, d\boldsymbol{\Theta}\ . \tag{15}$$

Sampling from (15) is straightforward: for each posterior sample $\{\boldsymbol{\Theta}^{(l)}\}$, $l = 1, 2, \ldots, L$, we draw $\boldsymbol{w}^{*(l)}$ from $p(\boldsymbol{w}^* \,|\, \boldsymbol{y}, \boldsymbol{\Theta}^{(l)})$ and then $\boldsymbol{w}_*(\boldsymbol{s}_0)^{(l)}$ from $p(\boldsymbol{w}_*(\boldsymbol{s}_0) \,|\, \boldsymbol{w}^{*(l)}, \boldsymbol{\Theta}^{(l)})$, which are both normal distributions. The procedure for Predictive process is analogous.

5.1. Model selection

To compare how well the different bias-adjustments perform, we adopt the posterior predictive loss approach of [32]. For each model, we simulate *independent* replicates for each observed outcome. Specifically, for the observed outcome $\boldsymbol{y}(\boldsymbol{s}_i)$ at location $\boldsymbol{s}_i$, we compute $p(\boldsymbol{y}_{rep}(\boldsymbol{s}_i) \,|\, \boldsymbol{y}) = \int p(\boldsymbol{y}_{rep}(\boldsymbol{s}_i) \,|\, \boldsymbol{\beta}, \boldsymbol{\Theta}) p(\boldsymbol{\Theta} \,|\, \boldsymbol{y}) d\boldsymbol{\Theta}$. Letting $\boldsymbol{\mu}_{rep,i}$ and $\boldsymbol{\Sigma}_{rep,i}$ be the posterior predictive mean and variance for each $\boldsymbol{y}_{rep}(\boldsymbol{s}_i)$, we will prefer models that will perform well under a decision-theoretic balanced loss function, penalizing both departure of replicated means from their observed values (lack of fit) and excessive uncertainty in the replicated data. Using a squared error loss function, the measures for these two criteria are evaluated as $G = \sum_{i=1}^{n} \|\boldsymbol{y}(\boldsymbol{s}_i) - \boldsymbol{\mu}_{rep,i}\|^2$, where $\|\cdot\|$ is the standard Euclidean norm and $P = \sum_{i=1}^{n} \mathrm{tr}(\boldsymbol{\Sigma}_{rep,i})$. We will use the score $D = G + P$ as a model selection criteria, with lower values of D indicating better models.

6. Illustrations

This section presents empirical performance of multivariate tapered predictive process along with its competitors through simulation studies and one real data application. In the simulation example, we analyze two datasets simulated from the classical Geostatistical model (4). The first dataset uses LMC with

exponential spatial correlation function for each $w_k(\boldsymbol{s})$ to simulate. For this simulation, we use PP, MPP and TPP with LMC and TPP with multivariate matern as the competitors. In the second simulation, we generate data from multivariate matern kernel with different degrees of smoothness for different components. For this simulation, PP, MPP, TPP with multivariate matern and TPP with LMC are used as competitors. Two simulations are used to indicate (a) how TPP under both specifications provide better inference than MPP and PP, (b) how the two different specifications of TPP work under misspecified model and (c) the impact on the performance depending on varying degree of smoothness in the individual components of the multivariate data generating process. All models are implemented in `R` with sparse matrix operations performed using the `SparseM` package in `R` . Details on `SparseM` can be found in `cran.r-project.org/web/packages/SparseM/SparseM.pdf`.

6.1. Analysis of synthetic data

Simulation 1

To illustrate the performance of the multivariate tapered predictive process (TPP), $2,000$ bivariate observations within a unit square domain are generated from the classical geostatistical model with likelihood $N(\boldsymbol{y} \,|\, \boldsymbol{\beta}\mathbf{1}_n, \boldsymbol{C}_w + \boldsymbol{I} \otimes \boldsymbol{\Psi})$ where $\boldsymbol{\Psi}$ is taken to be a diagonal matrix with two diagonal entries as ψ_1 and ψ_2. The mean of each location was assumed to have a common intercept denoted by $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. An exponential spatial correlation function was assumed for all spatial processes, i.e., θ_2 was fixed at 0.5 in (3). For the sake of identifiability of each element of $\boldsymbol{A}$, we assume $\boldsymbol{A} = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}$, with $A_{11} > 0$. The column labeled *True* in Table 1 depicts the mean and variance parameter values used to generate the data.

We fit TPP with both LMC and multivariate matern kernel for this data. As a competitor to TPP, multivariate predictive process and modified predictive process models with LMC are also fitted to the simulated data. For estimating

the LMC models, we assign a flat prior to the common intercept $\boldsymbol{\beta}$ and N(0,1) prior for A_{21}. All variance parameters (ψ_1, ψ_2, A_{11}, A_{22}) were assigned $IG(2,1)$ (mean = 1) prior. Decay parameters ϕ_1 and ϕ_2 in the exponential correlation functions for the two components of $\boldsymbol{w}(\boldsymbol{s})$ are assigned $U(1,10)$ prior that gives fairly wide support for range parameters given that the maximum inter-location distance in the generated data is 1.36. We remark that a uniform prior on the ϕ's does not translate into a uniform prior for the range, which, for the exponential correlation function, is customarily defined as $3/\phi$. Nevertheless, the prior range is broad enough to allow the data to drive the inference. For estimating TPP with multivariate matern, we assign a flat prior to the common intercept $\boldsymbol{\beta}$ and IG(2,1) prior for all the variance parameters. η_{12} is assigned $U(-0.9, 0.9)$ prior and θ is assigned a $U(1,10)$ prior that shows fairly broad range of prior support.

In estimating low rank models, we usually experiment with a varying number of knots. Typically, beyond a certain number of knots the substantive inference becomes robust. In this particular example, we found that placing just 50 knots randomly over the unit square domain was able to capture the salient features of the underlying true process. For brevity, we only present the analysis with 50 knots. In addition, while fitting tapered predictive process with LMC, we keep the taper range $\nu = .10$. For tapered predictive process with multivariate matern, we choose parameters $\nu_{k_1,k_2} = 0.10$, $k_1, k_2 = 1, 2$ so as to ensure the same taper range for the two tapered predictive process models.

Table 1 presents the posterior means along with 95% credible intervals for all the parameters. As expected, the global mean $\boldsymbol{\beta}$ is estimated robustly across all the models. Comparing TPP, MPP and PP with LMC, the nugget parameters ψ_1 and ψ_2 are found significantly overestimated by the predictive process model (PP), while the 95% CIs from the two bias adjusted models capture the true values of the nuggets. The spatial variance parameters (A_{11}, A_{22}) seem to be grossly overestimated in predictive process model, while the bias adjusted models tend to estimate them accurately. Note that TPP with multivariate matern is a misspecified model in this example, so that the spatial parameters are not comparable to the truth. The nugget parameters show a little overestimation

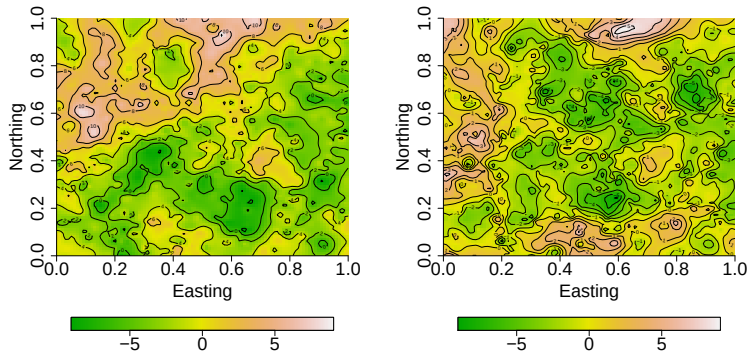
for this model.

Turning to model comparisons, we find the posterior predictive loss metric indicates improvements in performance for the two bias-adjusted models LMC models compared to the predictive process LMC model. Although TPP with multivariate matern is a mis-specified model, it exhibits model fit almost as good as MPP with LMC. Given the tapered covariance structure is "closest" (in the sense of Lemma 4.1) to the full Geostatistical model, it is also not very surprising that the tapered predictive process produces an overall better fit than the simple modified predictive process. A pictorial depiction of the residual spatial surfaces for outcomes 1 and 2 from MPP and TPP with LMC and TPP with multivariate matern can be found in Figure 2. Comparing these estimated residual surfaces with the true data generating surfaces in Figure 1 it is evident that among the two bias-adjusted processes, the tapered process is smoother. Needless to say, the residual surfaces constructed from the predictive process model are grossly oversmoothed (see Figure 1).

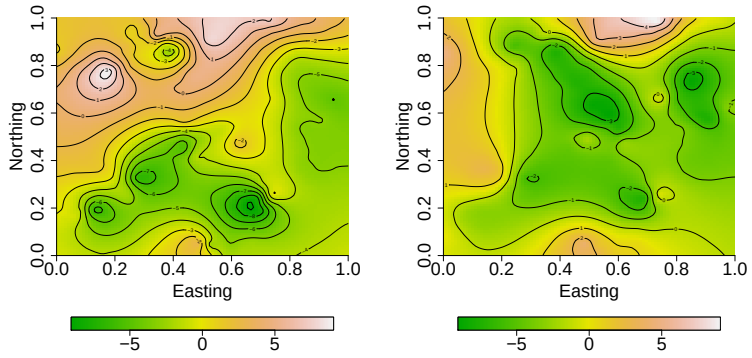
While comparing two bias adjusted models, our aim remains at accurately capturing the spatial correlation between any two observations. Figure 3 represents spatial correlations of 100 selected points from the full data set, constructed from the posterior estimates of modified, tapered and predictive process models for the two latent processes under LMC respectively. True exponential correlation curves are overlaid. It is evident that predictive process model proves to be the worst in capturing spatial correlation at small distances, though its performance begins to improve as distance increases. On the other hand, both modified predictive process and tapered adjusted models perform satisfactorily in capturing the true spatial correlation. Tapered predictive process model being particularly notable as the spatial correlations from this model closely coincide with the true exponential correlation (see Figure 3) from the full model. These plots provide convincing evidence regarding the superior performance of tapered predictive process with respect to its competitors.

Simulation 2

Simulation 2 is conducted to understand how the models behave when the bi-

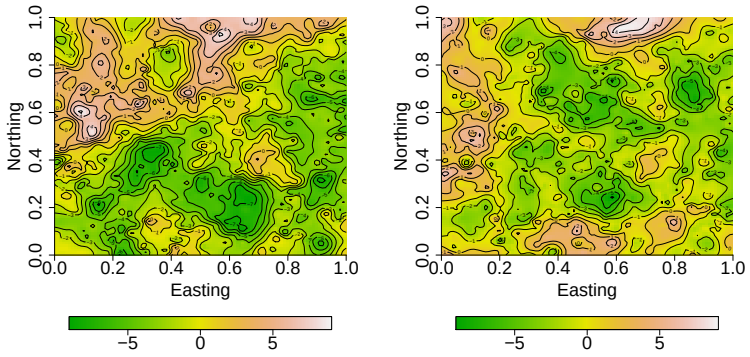


(a) $w_1^{(1)}(\boldsymbol{s})$ (b) $w_2^{(1)}(\boldsymbol{s})$

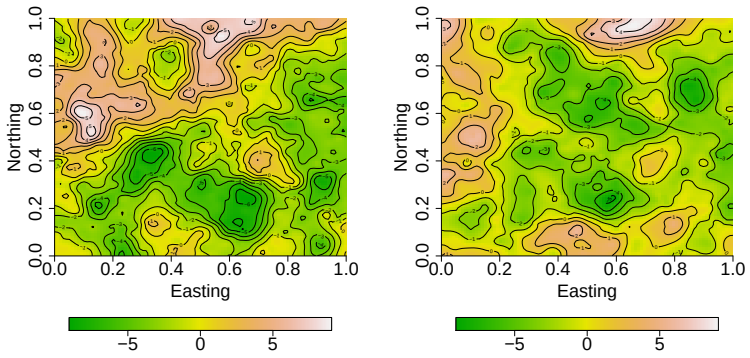


(c) $w_{pp,1}^{(1)}(\boldsymbol{s})$ (d) $w_{pp,2}^{(1)}(\boldsymbol{s})$

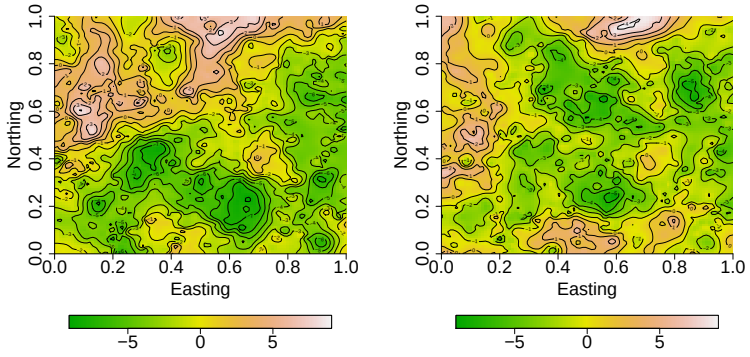
Figure 1: 1(a) & 1(b) provide true spatial surfaces. 1(c) & 1(d) provide estimated mean spatial surfaces for predictive process with LMC.



(a) $w^{(1)}_{tap,1}(\boldsymbol{s})$ (b) $w^{(1)}_{tap,2}(\boldsymbol{s})$

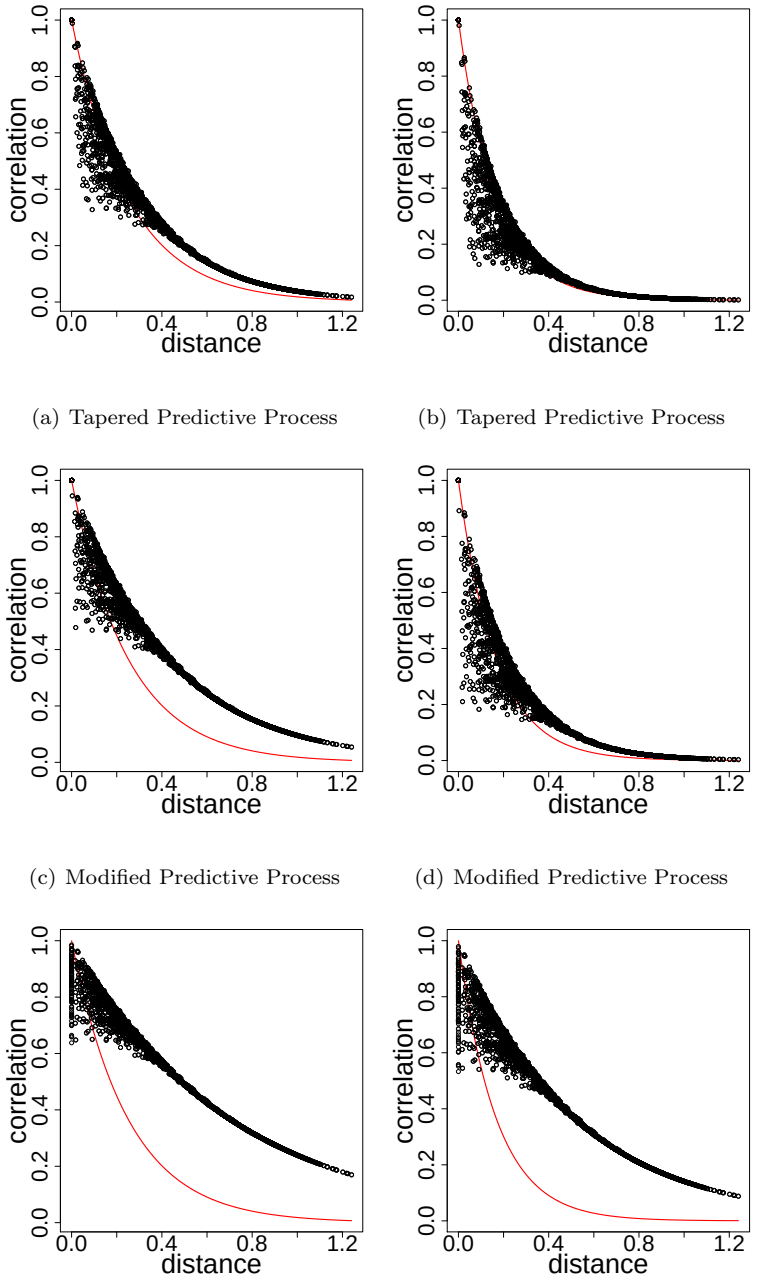


(c) $w^{(2)}_{tap,1}(\boldsymbol{s})$ (d) $w^{(2)}_{tap,2}(\boldsymbol{s})$



(e) $w^{(1)}_{mpp,1}(\boldsymbol{s})$ (f) $w^{(1)}_{mpp,2}(\boldsymbol{s})$

Figure 2: Column 1 gives estimated spatial surfaces (posterior mean) for the first spatial component of TPP with LMC, TPP with multivariate matern & MPP with LMC. Column 2 provides estimated spatial surfaces for the second spatial component of the same.



(a) Tapered Predictive Process

(b) Tapered Predictive Process

(c) Modified Predictive Process

(d) Modified Predictive Process

(e) Predictive Process

(f) Predictive Process

Figure 3: Estimated spatial correlation for the latent process, Column 1: $v_1(\boldsymbol{s})$ overlaid with the true exponential correlation. Column 2: $v_2(\boldsymbol{s})$ overlaid with the true exponential correlation.

Table 1: The median and 95% Bayesian credible intervals of parameters for three spatial models – the predictive process model (PP), modified predictive process (MPP) and tapered predictive process (TPP) – are presented for the synthetic data set. TPP(LMC) and TPP(MM) show results for the tapered predictive process under linear model co-regionalization and multivariate matern. Also presented are model comparison metrics.

	True	PP(LMC)	MPP(LMC)	TPP(LMC)	TPP(MM)
β_0	5	6.94 (3.21 , 11.37)	6.34 (4.05 , 9.44)	5.92 (4.17 , 7.97)	4.99 (3.61 , 6.51)
β_1	1	-0.82 (-4.83, 1.47)	0.48 (-0.60,1.68)	0.48 (-0.38, 1.48)	0.20 (-0.84, 1.41)
ψ_1	.5	2.42 (2.22, 2.67)	0.66 (0.38, 0.94)	0.44 (0.32, 0.57)	0.71 (0.56, 1)
ψ_2	.4	1.55 (1.44, 1.69)	0.44 (0.24, 0.67)	0.31 (0.23, 0.42)	0.52 (0.44, 0.63)
A_{11}	3	6.63 (4.78, 8.98)	3.04 (2.50, 4.32)	2.68 (2.28, 3.37)	–
A_{12}	.9	1.67 (0.10, 3.15)	1.03 (0.78, 1.41)	0.81 (0.63, 1.05)	–
A_{22}	2	3.46 (2.14, 5.25)	1.62 (1.40, 2.01)	1.60 (1.42, 1.92)	–
ϕ_1	4	1.43 (1.02 , 2.32)	2.35 (1.12 , 3.51)	3.27 (1.95 , 4.84)	–
ϕ_2	6	1.96 (1.05 , 4.77)	4.66 (2.55 , 6.60)	5.47 (3.29 , 7.28)	–
G	–	3906.87	361.72	214.72	653.14
P	–	4221.81	2077.43	1501.29	2147.02
D	–	8128.68	2439.15	1716.02	2800.16

variate spatial random effects are generated from bivariate matern correlation (described in Section 2.1.2) with component 1 and 2 having smoothness parameters 0.5 and 1.5 respectively, i.e. the two components having two different degrees of smoothness. Table 2 depicts the mean and variance parameter values used to generate the data.

We fit TPP with both LMC and multivariate matern kernel for this data. As a competitor to TPP, multivariate modified predictive process model with multivariate matern kernel is fitted to the simulated data. Prior specifications for all parameters remains the same as simulation 1. Similar to simulation 1, we employ the tapering correlation kernel outlined in **Strategy 1** for multivariate TPP with LMC. For its implementation, we take $\nu = 0.10$. For multivariate TPP with multivariate matern kernel outlined in **Strategy 2**, we choose $\nu_{k_1,k_2} =$

0.10, $k_1, k_2 = 1 : m$, so as to keep the same taper range for the two competing approaches. In general, we fix $\nu_{k_1,k_2} = \nu$, vary ν (not shown here) for the competing procedures and obtain similar relative performances for both of them. Additionally, we found that r_{12} in **Strategy 2** has an effect on the inference which is discussed as we proceed. Multivariate predictive process is ignored as a competitor as it has already demonstrated inferior performance in simulation 1.

Table 2 presents the posterior means along with 95% credible intervals for all the parameters. Note that the result presented here assumes $r_{12} = 1$ which produces best performance. The sensitivity of inference to the choice of r_{12} is discussed later. As expected, the global mean $\boldsymbol{\beta}$ is estimated robustly across all the models. The 95% CIs of the nugget parameters ψ_1, ψ_2 from the TPP models capture the true values of the nuggets, while MPP with multivariate matern shows little underestimation. The estimates of the spatial parameters are found to be a bit off from the true values, perhaps they are very weakly identifiable. Note that TPP with LMC is a misspecified model in this example, so that the spatial parameters are not comparable to the truth.

In terms of posterior predictive loss criterion, TPP with multivariate matern exhibits superior performance. Similar to simulation 1, tapered predictive process produces an overall better fit than the simple modified predictive process. A pictorial depiction of the residual spatial surfaces for outcomes 1 and 2 from MPP with multivariate matern and TPP with LMC and TPP with multivariate matern can be found in Figure 5. Comparing these estimated residual surfaces with the true data generating surfaces in Figure 4 it is evident that among the two bias-adjusted processes, the tapered process is smoother. More importantly, TPP with multivariate matern is able to capture two surface with varying degrees of smoothness more accurately than its competitors.

As in simulation 1, we also plot the estimated posterior spatial correlations of $w^{(2)}_{tpp,1}$ and $w^{(2)}_{tpp,2}$ and overlay the true data generating Matern correlation function on them. It is pretty clear that both short range and long range correlations are satisfactorily estimated by the tapered predictive process under

Table 2: The median and 95% Bayesian credible intervals of parameters for three spatial models – the predictive process model (PP), modified predictive process (MPP) and tapered predictive process (TPP) – are presented for the synthetic data set. TPP(LMC) and TPP(MM) show results for the tapered predictive process under linear model co-regionalization and multivariate matern. Also presented are model comparison metrics.

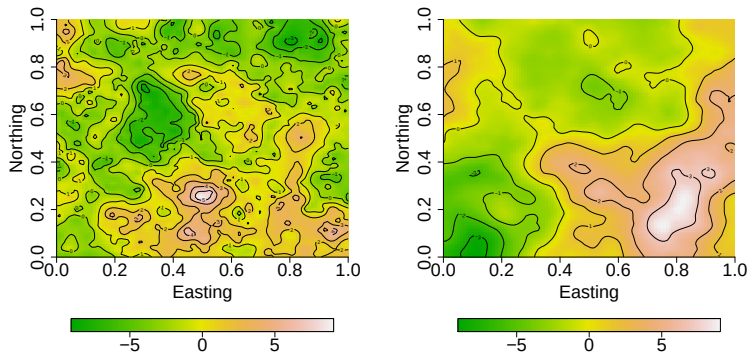
	True	MPP(MM)	TPP(MM)	TPP(LMC)
β_0	4	4.40 (3.77 , 5.03)	4.50 (3.92 , 5.04)	4.43 (3.00 , 5.35)
β_1	-1	-1.69 (-3.02,-0.41)	-0.76 (-1.71, 0.13)	-0.77 (-2.46, 0.27)
ψ_1	.2	0.17 (0.15, 0.19)	0.24 (0.19, 0.37)	0.23 (0.18,0.29)
ψ_2	.1	0.09 (0.08, 0.10)	0.09 (0.08, 0.10)	0.09 (0.08, 0.10)
θ	8	3.62 (3.28, 3.92)	5.57 (5.09, 6.34)	–
σ_1^2	1.7	0.88 (0.60, 1.19)	1.36 (1.16, 1.68)	–
σ_2^2	1.5	1.75 (1.42, 2.25)	1.27 (0.90, 1.62)	–
η_{12}	0.2	0.16 (0.12 , 0.21)	0.06 (0.06 , 0.06)	–
G	–	254.41	195.14	256.12
P	–	1585.72	613.37	616.87
D	–	1840.13	808.51	872.99

multivariate Matern.

Finally, we compare the model fitting statistics of TPP with multivariate matern with different values of $r_{12} = 0.2, 0.5, 1$. Table 3 shows a considerable improvement as r_{12} increases, with the best performance shown with $r_{12} = 1$. This is consistent with the findings of [26] where the authors argue choosing for $r_{12} = 1$ in the tapering kernel. Our investigations also reveal the best model fitting is shown by $r_{12} = 1$. Therefore, we proceed with this choice of r_{12} for the data analysis.

6.2. Forestry example

Basal area (BAREA) of the tree together with volume (VOL) are used, in conjunction with other information, to assess a tree's economic and ecological value. They are frequently used in examining a forest's productivity and growth rate, thereby playing an important role in local, regional and global scale deci-



(a) $w_1^{(2)}(\boldsymbol{s})$ (b) $w_2^{(2)}(\boldsymbol{s})$

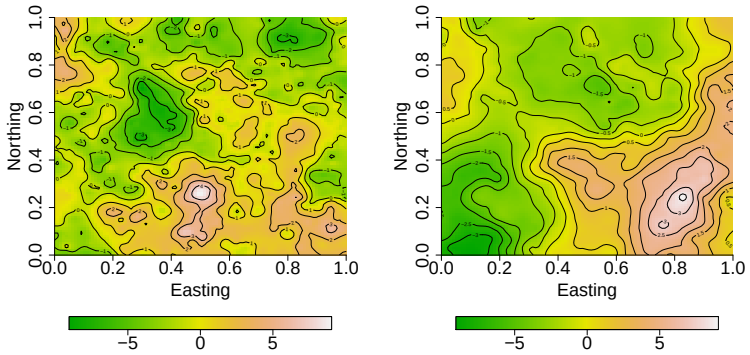
Figure 4: 4(a) & 4(b) provide true spatial surfaces.

Table 3: Comparison between TPP with multivariate matern for $r_{12} = 0.2, 0.5, 1$.

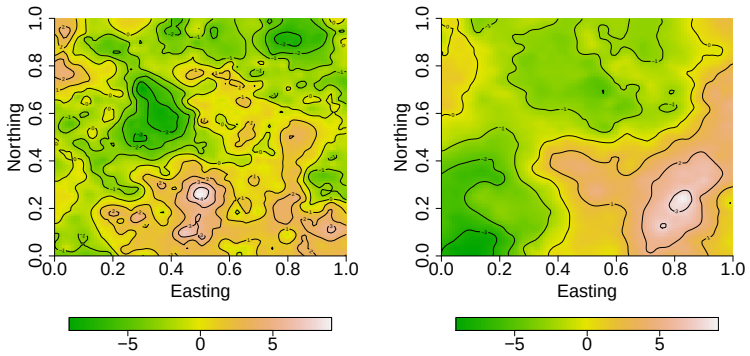
r_{12}	0.2	0.5	1
G	279.34	206.06	195.14
P	722.12	635.91	613.37
D	1001.46	841.97	808.51

sions. Basal area is the cross sectional area of a tree at the diameter at breast height (DBH; 1.37m above the forest floor). Measuring VOL is more cumbersome and expensive as compared to measuring Basal area (BAREA). Therefore VOL is often measured on a small subset of those trees that also provide BAREA measurements. Then a statistical model that relates VOL to BAREA is used to predict VOL for the complement of this subset.

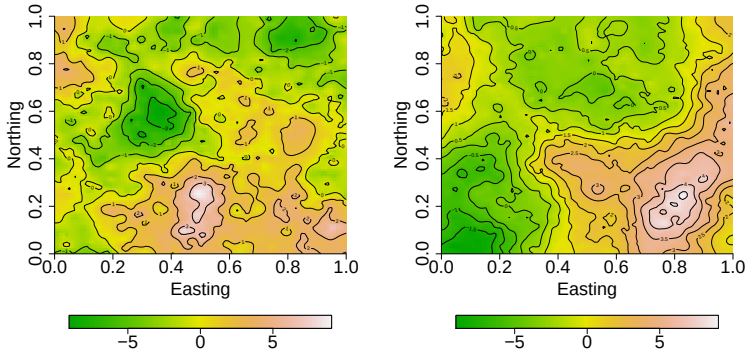
The dataset we analyze is collected from an inventory conducted in parts of the Zurichberg Forest belonging to the city and Canton of Zurich (see [33]). The inventoried area covered 217.9 ha, of which 17.1 ha served for the full census, in which the tree coordinates have been recorded. The inventory utilizes systematic cluster sampling schemes where cluster comprises five points: *central point, two points established 30m east or west of the central point; two other points each*



(a) $w^{(1)}_{tap,1}(\boldsymbol{s})$ (b) $w^{(1)}_{tap,2}(\boldsymbol{s})$

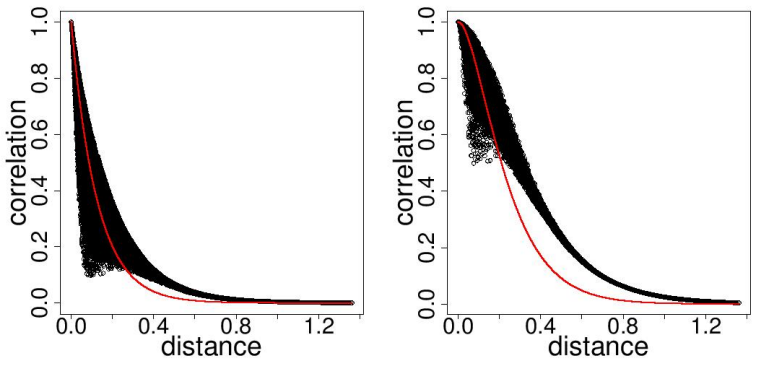


(c) $w^{(2)}_{tap,1}(\boldsymbol{s})$ (d) $w^{(2)}_{tap,2}(\boldsymbol{s})$



(e) $w^{(2)}_{mpp,1}(\boldsymbol{s})$ (f) $w^{(2)}_{mpp,2}(\boldsymbol{s})$

Figure 5: Column 1 gives estimated spatial surfaces (posterior mean) for the first spatial component of TPP with LMC, TPP with multivariate matern & MPP with multivariate matern. Column 2 provides estimated spatial surfaces for the second spatial component of the same.



(a) TPP(MM):first component (b) TPP(MM):second component

Figure 6: Estimated posterior spatial correlation for the two latent processes in $\boldsymbol{w}_{tap}^{(2)}$. True data generating correlation functions are overlaid in red curve.

established 40m either north or south of the the cental point. At these five points, a number of measurements, including BAREA and VOL, have been collected for 4954 trees. Given this data set, a forestry scientist is always interested in predicting the volume (VOL) of a tree based on the Basal area (BAREA) measurement of the tree at some location.

We propose to use a spatial analysis based on bivariate vector, with individual entries as VOL and BAREA, at each location. Our candidate models include predictive process, modified predictive process and tapered predictive process with LMC and tapered predictive process with the multivariate matern kernel. It is to be noted that from bivariate spatial model, univariate conditionals (in particular VOL given BAREA) can readily be obtained.

The locations are all mapped to a $[0, 1] \times [0, 1]$ domain. Like in the simulation experiment, we present, again for brevity, the results with only 50 knots, distributed uniformly over the domain. The priors for ψ_1, ψ_2 and A_{11}, A_{22} follow $IG(2\ \ 1)$. We assume exponential spatial correlation functions for the two bivariate processes, and assign $U(1, 10)$ prior to both ϕ's, which corresponds to broad ranges of support given the maximum distance between any two trees

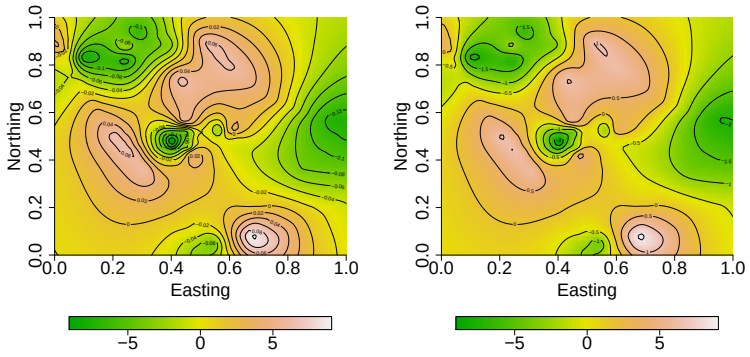
is 1.22 (in the transformed scale). For all models the intercepts are assigned completely *flat* noninformative priors. For tapered predictive process with multivariate matern kernel, the priors similar to *Simulation 1* and *Simulation 2* are used. Finally, both for TPP with multivariate matern and LMC, we use the same taper range and tapering kernels as in *Simulation 1* and *Simulation 2*.

Table 4: The median and 95% Bayesian credible intervals for the four spatial models – the predictive process model with LMC, the modified predictive process with LMC and the tapered predictive process with LMC and multivariate matern – are presented for the forestry example. They are abbreviated as PP(LMC), MPP(LMC), TPP(LMC) and TPP(MM). Also presented are model comparison metrics.

	PP(LMC)	MPP(LMC)	TPP(LMC)	TPP(MM)
β_0	0.15 (0.01,0.30)	0.14 (0.10 , 0.18)	0.17 (0.09 , 0.24)	0.22 (-0.32, 0.82)
β_1	2.19 (0.15, 4.49)	2.14 (1.40,2.84)	2.57 (1.36, 3.79)	2.62 (-1.14, 7.28)
ψ_1	0.01 (0.009,0.01)	0.001 (0.001, 0.001)	0.001 (0.001, 0.001)	0.007 (0.007,0.008)
ψ_2	2.25 (2.08, 2.45)	0.02 (0.02, 0.03)	0.03 (0.02, 0.03)	0.02 (0.02,0.02)
G	2795.94	8.98	6.88	12.68
P	2818.62	485.61	444.16	765.43
D	5614.56	494.59	451.04	778.11

The inference is based upon running three parallel chains for 5000 iterations each and discarding the first 2000 iterations as pre-convergence burn-in. Since only β_0, β_1, ψ_1, ψ_2 and G,P, D scores are comparable across models, they are only presented in Table 4. As seen in the synthetic example, the estimates for ψ_1 and ψ_2 from bias-adjusted models are more reliable than those from the predictive process model. The G,P,D scores again reveal the considerable improvements in overall model fit achieved by the bias-adjusted models over the predictive process model.

Figure 7 and 8 display the posterior means of the residual surface from different models. The two figures reveal excessive oversmoothing by the predictive process, while the bias-adjustments compensate for such excesses. Of the two



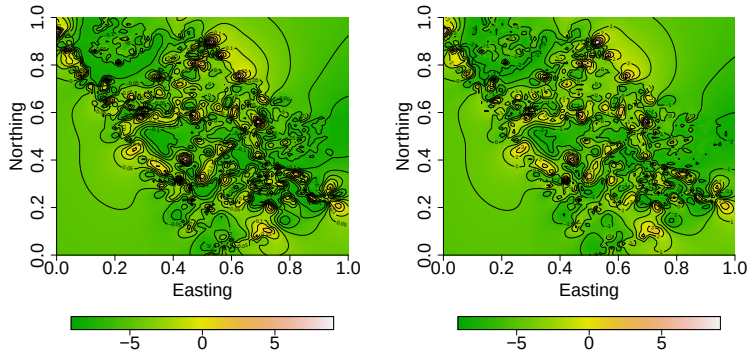
(a) $w^{(1)}_{pp,1}$ (b) $w^{(1)}_{pp,2}$

Figure 7: Column 1 shows the estimated (posterior mean) residual spatial surface of BAREA from multivariate predictive process LMC. The estimated spatial surface of VOL from multivariate predictive process LMC is in column 2.

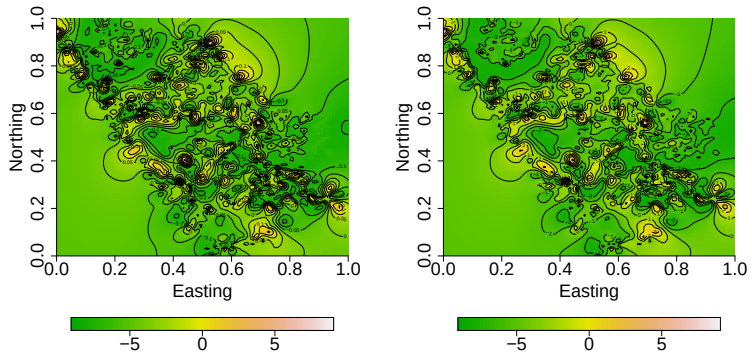
bias-adjusted models, the tapered process under LMC provides marginally better fit than MPP under LMC. This is consistent with our observations in the simulation study. TPP with multivariate matern also shows fit closely comparable to TPP with LMC and MPP with LMC.

7. Conclusion and Further work

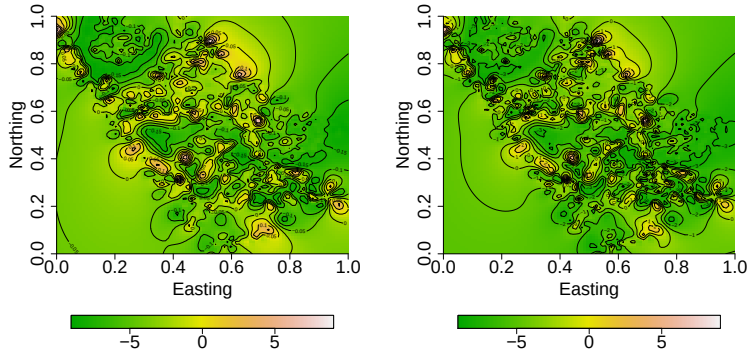
This manuscript formally explores the nature of biases in residual variability captured by multivariate low-rank spatial (geostatistical) models. We have explained how such biases arise and show, specifically, how low-rank multivariate predictive processes can help quantify such biases. We have then proceeded to formulate a multivariate "bias-adjusted" tapered predictive process model that attempts to ameliorate the impact of this bias and lead to improved model performance. We proposed a new class of multivariate tapered predictive processes based on the multivariate matern kernel. We compared its performance with multivariate tapered predictive processes based on LMC. We have proved new results on the smoothness of spatial processes and argue the usage of tapered



(a) $w^{(1)}_{mpp,1}$ (b) $w^{(1)}_{mpp,2}$



(c) $w^{(1)}_{tap,1}$ (d) $w^{(1)}_{tap,2}$



(e) $w^{(2)}_{tap,1}$ (f) $w^{(2)}_{tap,2}$

Figure 8: Column 1 shows the estimated (posterior mean) residual spatial surfaces of BAREA from the three candidate models, MPP(LMC), TPP(LMC) and TPP(MM). The estimated spatial surfaces of VOL from the three candidate models are in column 2.

predictive process as a computationally convenient alternative to Gaussian process (GP) to study abrupt changes in the spatial surface.

As a future work, we propose to extend multivariate tapered predictive process to account for the space varying correlation between components. It is also of great importance to implement directional tapering as opposed to the distance based tapering to take care of the long range correlations in a few directions.

Appendix A

Proof of proposition 4.1

Each of the dispersion metrics in Section 4.1 can be expressed as

$$\Delta_{(k)}^{(1)} = \|\mathcal{A}[\mathcal{C}_v' \boldsymbol{C}_v^{*-1} \mathcal{C}_v + \boldsymbol{D}_{(k),1} - \boldsymbol{C}_v]\mathcal{A}'\|_2 \ , k = 0, 1, 2$$

where $\boldsymbol{D}_{(0),1}$ is the zero matrix so $\Delta_{(0)}$ corresponds to the predictive process and $k = 1, 2$ correspond to the modified predictive process and the tapered adjustments respectively.

Let $\boldsymbol{H}_{ij}$ denote the $m \times m$ $(i,\ j)$-th block element of $(\boldsymbol{C}_v - \mathcal{C}_v' \boldsymbol{C}_v^{*-1} \mathcal{C}_v)$; $\boldsymbol{H}_{ij}$'s remain invariant to the four models. Let $\boldsymbol{B}_{ij}^{(k)}$ denote the elements of $\boldsymbol{C}_v - \mathcal{C}_v' \boldsymbol{C}_v^{*-1} \mathcal{C}_v - \boldsymbol{D}_{(k)}$, and let $\boldsymbol{T}_{ij}$ be the block elements of $\boldsymbol{T}$ (for the tapered model). $\boldsymbol{H}_{ij}, \boldsymbol{T}_{ij}$ are diagonal. Note Then,

$$\boldsymbol{B}_{ij}^{(1)} = \boldsymbol{H}_{ij}, \quad \text{if} \quad i \neq j \ ; \quad \boldsymbol{B}_{ii}^{(1)} = \boldsymbol{0} \tag{16}$$

$$\boldsymbol{B}_{ij}^{(2)} = \boldsymbol{H}_{ij} \odot (\boldsymbol{I} - \boldsymbol{T}_{ij}), \quad \text{if} \quad i \neq j \ , \quad \boldsymbol{B}_{ii}^{(2)} = 0 \tag{17}$$

It is quite straightforward to see,

$$\boldsymbol{B}_{ij}^{(2)} \preceq \boldsymbol{B}_{ij}^{(1)} \preceq \boldsymbol{H}_{ij}, \tag{18}$$

where, $\boldsymbol{A}_1 \preceq \boldsymbol{A}_2$ means $\boldsymbol{A}_2$ is elementwise greater than or equal to $\boldsymbol{A}_1$.

We will state the following lemma, the proof of which follows from induction method.

Lemma 7.1. *Let* $\boldsymbol{S} = diag\,(S_1, ..., S_m)$, *then* $\|\boldsymbol{ASA'}\|_2^2 = \sum_{l=1}^{m} S_l^2[\sum_{k=l}^{m} a_{kl}^2]^2 + 2\sum_{l<l'} S_l S_{l'}[\sum_{k=l'}^{m} a_{kl}a_{kl'}]^2$

Let the l-th diagonal element of $\boldsymbol{B}_{ij}^{(1)}$, $\boldsymbol{B}_{ij}^{(2)}$ and $\boldsymbol{H}_{ij}$ be given by $b_{ijl}^{(1)}$, $b_{ijl}^{(2)}$ and h_{ijl} respectively. Applying the aforementioned lemma,

$$\Delta_{(2)} = \left(\sum_{i,j=1}^{m}\left\{\sum_{l=1}^{m} b_{ijl}^{(2)2}\left[\sum_{k=l}^{m} a_{kl}^2\right]^2 + 2\sum_{l<l'} b_{ijl}^{(2)} b_{ijl'}^{(2)}\left[\sum_{k=l'}^{m} a_{kl}a_{kl'}\right]^2\right\}\right)^{\frac{1}{2}}$$

$$\Delta_{(1)} = \left(\sum_{i,j=1}^{m}\left\{\sum_{l=1}^{m} b_{ijl}^{(1)2}\left[\sum_{k=l}^{m} a_{kl}^2\right]^2 + 2\sum_{l<l'} b_{ijl}^{(1)} b_{ijl'}^{(1)}\left[\sum_{k=l'}^{m} a_{kl}a_{kl'}\right]^2\right\}\right)^{\frac{1}{2}}$$

$$\Delta_{(0)} = \left(\sum_{i,j=1}^{m}\left\{\sum_{l=1}^{m} h_{ijl}^{2}\left[\sum_{k=l}^{m} a_{kl}^2\right]^2 + 2\sum_{l<l'} h_{ijl} h_{ijl'}\left[\sum_{k=l'}^{m} a_{kl}a_{kl'}\right]^2\right\}\right)^{\frac{1}{2}}$$

The rest follows from (18).

proof of Theorem 4.3

Before proving Theorem 4.3 we will prove the following Proposition,

Proposition 1

If $Z_1(\boldsymbol{s})$ and $Z_2(\boldsymbol{s})$ are two different mean square differentiable Gaussian process then,

1. $Z_1(\boldsymbol{s}) + Z_2(\boldsymbol{s})$ and $Z_1(\boldsymbol{s}) - Z_2(\boldsymbol{s})$ are mean square differentiable.
2. If $Z_1(\boldsymbol{s})$ and $Z_2(\boldsymbol{s})$ are independent processes then $Z_1(\boldsymbol{s})Z_2(\boldsymbol{s})$ is also mean square differentiable.

proof: for any $\boldsymbol{s}$, $Z_1(\cdot)$ and $Z_2(\cdot)$ are differentiable at $\boldsymbol{s}$ means, $\exists$ functions ∇Z_1 and ∇Z_2 respectively, s.t., for any vector $\boldsymbol{u}$ with $||\boldsymbol{u}|| = 1$ we have,

$$\lim_{h\to 0} \mathrm{E}\left[\frac{Z_1(\boldsymbol{s}+h\boldsymbol{u}) - Z_1(\boldsymbol{s})}{h} - \langle\nabla Z_1(\boldsymbol{s}), \boldsymbol{u}\rangle\right]^2 = 0 \qquad (19)$$

$$\lim_{h\to 0} \mathrm{E}\left[\frac{Z_2(\boldsymbol{s}+h\boldsymbol{u}) - Z_2(\boldsymbol{s})}{h} - \langle\nabla Z_2(\boldsymbol{s}), \boldsymbol{u}\rangle\right]^2 = 0 \qquad (20)$$

Let $X_h^{(1)} = \left\{ \frac{Z_1(\boldsymbol{s}+h\boldsymbol{u})-Z_1(\boldsymbol{s})}{h} - \langle \nabla Z_1(\boldsymbol{s}), \boldsymbol{u} \rangle \right\}$, $X_h^{(2)} = \left\{ \frac{Z_2(\boldsymbol{s}+h\boldsymbol{u})-Z_2(\boldsymbol{s})}{h} - \langle \nabla Z_2(\boldsymbol{s}), \boldsymbol{u} \rangle \right\}$ and $X_h^{(3)} = \{Z_1(\boldsymbol{s}+h\boldsymbol{u}) - Z_1(\boldsymbol{s})\}$. (1) follows from the fact,

$$\begin{aligned} & \lim_{h\to 0} \quad \mathrm{E}\left[\frac{(Z_1 \overset{+}{-} Z_2)(\boldsymbol{s}+h\boldsymbol{u}) - (Z_1 \overset{+}{-} Z_2)(\boldsymbol{s})}{h} - \langle \nabla Z_1(\boldsymbol{s}) \overset{+}{-} \nabla Z_2(\boldsymbol{s}), \boldsymbol{u} \rangle \right]^2 \\ \leq & \lim_{h\to 0} \quad 2\left\{ \mathrm{E}\left[X_h^{(1)}\right]^2 + E\left[X_h^{(2)}\right]^2 \right\} \\ = & \quad 0 \end{aligned}$$

(2) Now assume $Z_1(\boldsymbol{s})$ and $Z_2(\boldsymbol{s})$ are independent.

$$\begin{aligned} & \lim_{h\to 0} \quad \mathrm{E}\left[\frac{(Z_1Z_2)(\boldsymbol{s}+h\boldsymbol{u}) - (Z_1Z_2)(\boldsymbol{s})}{h} - \langle Z_2(\boldsymbol{s})\nabla Z_1(\boldsymbol{s}) + Z_1(\boldsymbol{s})\nabla Z_2(\boldsymbol{s}), \boldsymbol{u} \rangle \right]^2 \\ = & \lim_{h\to 0} \quad \mathrm{E}\left[Z_1(\boldsymbol{s}+h\boldsymbol{u})X_h^{(2)} + Z_2(\boldsymbol{s})X_h^{(1)} + X_h^{(3)}\langle \nabla Z_2(\boldsymbol{s}), \boldsymbol{u} \rangle \right]^2 \\ \leq & \lim_{h\to 0} \quad \left\{ \left(\mathrm{E}\left[\boldsymbol{Z}_1(\boldsymbol{s}+h\boldsymbol{u})\boldsymbol{X}_h^{(2)} \right]^2 \right)^{\frac{1}{2}} + \left(\mathrm{E}\left[\boldsymbol{Z}_2(\boldsymbol{s})\boldsymbol{X}_h^{(1)} \right]^2 \right)^{\frac{1}{2}} + \left(\mathrm{E}\left[\boldsymbol{X}_h^{(3)}\langle \nabla \boldsymbol{Z}_2(\boldsymbol{s}), \boldsymbol{u} \rangle \right]^2 \right)^{\frac{1}{2}} \right\}^2 \\ = & \quad 0 \end{aligned}$$

Last two steps follow by direct application of *Minkowski Inequality* and independence of $\boldsymbol{Z}_1(\boldsymbol{s})$ and $\boldsymbol{Z}_2(\boldsymbol{s})$.

Proof of lemma 4.3

- We know, the function $H_j(\boldsymbol{s}) = \frac{\sigma^2}{2^{\theta_2-1}\Gamma(\theta_2)}(||\boldsymbol{s}-\boldsymbol{s}_j^*||\theta_1)^{\theta_2}\kappa_{\theta_2}(||\boldsymbol{s}-\boldsymbol{s}_j^*||;\theta_1)$ is totally differentiable except at $\boldsymbol{s}_j^*$. Therefore $\exists \nabla \boldsymbol{H}_j(\boldsymbol{s}) = (\nabla H_{j_1}(\boldsymbol{s}), \nabla H_{j_2}(\boldsymbol{s}))'$ s.t. for any vector $\boldsymbol{u}$ with $||\boldsymbol{u}|| = 1$ we have,

$$H_j(\boldsymbol{s}+h\boldsymbol{u}) = H_j(\boldsymbol{s}) + h\nabla \boldsymbol{H}_j'(\boldsymbol{s})\boldsymbol{u} + o(h), \forall j = 1, 2, ..., n^*, \boldsymbol{s} \in \mathcal{R}^2 - \mathcal{S}^* \quad (21)$$

therefore,

$$\lim_{h\to 0} X_j(\boldsymbol{s}, h) = \lim_{h\to 0} \frac{H_j(\boldsymbol{s}+h\boldsymbol{u}) - H_j(\boldsymbol{s})}{h} - \nabla \boldsymbol{H}_j'(\boldsymbol{s})\boldsymbol{u} = 0. \ \boldsymbol{s} \in \mathcal{R}^2 - \mathcal{S}^* \quad (22)$$

Let, $\nabla \boldsymbol{H} = (\nabla \boldsymbol{H}_1(\boldsymbol{s}), \ldots, \nabla \boldsymbol{H}_{n^*}(\boldsymbol{s}))$, then,

$$\begin{aligned}
&\lim_{h\to 0} \quad \mathrm{E}\left[\frac{w_{pp}(\boldsymbol{s}+h\boldsymbol{u}) - w_{pp}(\boldsymbol{s})}{h} - \langle \nabla \boldsymbol{H} \boldsymbol{C}_v^{*-1}\boldsymbol{w}^*, \boldsymbol{u}\rangle\right]^2 \\
&= \quad \lim_{h\to 0} \mathrm{E}\left[\left\{\frac{\boldsymbol{c}_v(\boldsymbol{s}+h\boldsymbol{u})' - \boldsymbol{c}_v(\boldsymbol{s})'}{h} - \boldsymbol{u}'\nabla \boldsymbol{H}\right\}\boldsymbol{C}_v^{*-1}\boldsymbol{w}^*\right]^2 \\
&= \quad \lim_{h\to 0}\{X_1(\boldsymbol{s},h), \ldots, X_{n^*}(\boldsymbol{s},h)\}'\boldsymbol{C}_v^{*-1}\{X_1(\boldsymbol{s},h), \ldots, X_{n^*}(\boldsymbol{s},h)\} \\
&= \quad 0.
\end{aligned}$$

The last equality follows from (22) and the fact that $\boldsymbol{C}_v^{*-1}$ doesn't involve h. Therefore, with an appeal to the arbitrariness of $\boldsymbol{s}$ and $\boldsymbol{u}$, we conclude that Predictive Process is L_2 differentiable.

Let, $\nabla \boldsymbol{H}_j(\boldsymbol{s}) = (\nabla \boldsymbol{H}_{j_1}(\boldsymbol{s}), \nabla \boldsymbol{H}_{j_2}(\boldsymbol{s}))'$ with $\nabla \boldsymbol{H}_{j_i}(\boldsymbol{s})$ being totally differentiable except at the set of knot points. Continuing in this way, we define,

$$\nabla^l \boldsymbol{H}_{j_{i_1,i_2,\ldots,i_{l-1}}}(\boldsymbol{s}) = (\nabla^l \boldsymbol{H}_{j_{i_1,i_2,\ldots,i_{l-1},1}}(\boldsymbol{s}), \nabla^l \boldsymbol{H}_{j_{i_1,i_2,\ldots,i_{l-1},d}}(\boldsymbol{s}))$$

for any l and any $(i_1, ..., i_{l-1}) \in \{1, 2\}^{l-1}$

Note that, in order to prove the fact that Predictive process is infinitely differentiable, it is enough to prove that Predictive process is differentiable for any $k_1 \geq 1$. Let, it be already k_1-times mean square differentiable. We will show, it is $(k_1 + 1)$-times mean square differentiable.

It is not difficult to see that $\nabla^{k_1}\tilde{w}(\boldsymbol{s}) = \nabla^{k_1}\boldsymbol{H}(\boldsymbol{s})\boldsymbol{C}_v^{*-1}\boldsymbol{w}^*$, where, $\nabla^{k_1}\boldsymbol{H}(\boldsymbol{s})$ is a $2^{k_1} \times n^*$ matrix with a typical row is of the form

$$(\nabla^{k_1}\boldsymbol{H})_{i_1,\ldots,i_{k_1}}(\boldsymbol{s}) = (\nabla^{k_1}\boldsymbol{H}_{1_{i_1,i_2,\ldots,i_{k_1}}}(\boldsymbol{s}), ..., \nabla^{k_1}\boldsymbol{H}_{n^*{}_{i_1,\ldots,i_{k_1}}}(\boldsymbol{s})).$$

It is now enough to show that, $(\nabla^{k_1}\boldsymbol{H})_{i_1,\ldots,i_{k_1}}(\boldsymbol{s})\boldsymbol{C}_v^{*-1}\boldsymbol{w}^*$ is mean square differentiable.

$$\begin{aligned}
\lim_{h\to 0} X_{j_{i_1,\ldots,i_{k_1}}}(\boldsymbol{s},h) &= \lim_{h\to 0}\frac{\nabla^{k_1}\boldsymbol{H}_{j_{i_1,i_2,\ldots,i_{k_1}}}(\boldsymbol{s}+h\boldsymbol{u}) - \nabla^{k_1}\boldsymbol{H}_{j_{i_1,i_2,\ldots,i_{k_1}}}(\boldsymbol{s})}{h} \\
&- \quad \nabla^{k_1+1}\boldsymbol{H}_{j_{i_1,i_2,\ldots,i_{k_1}}}(\boldsymbol{s})'\boldsymbol{u} = 0.\ \boldsymbol{s} \in \mathcal{R}^2 - \mathcal{S}^* \qquad (23)
\end{aligned}$$

Now,

$$\mathrm{E}\left[\left\{\frac{\nabla^{k_1}\boldsymbol{H}_{i_1,i_2,\ldots,i_{k_1}}(\boldsymbol{s}+h\boldsymbol{u})-\nabla^{k_1}\boldsymbol{H}_{i_1,i_2,\ldots,i_{k_1}}(\boldsymbol{s})}{h}-\nabla^{k_1+1}\boldsymbol{H}_{i_1,i_2,\ldots,i_{k_1}}(\boldsymbol{s})\right\}\boldsymbol{C}_v^{*-1}\boldsymbol{w}^*\right]^2$$
$$=\left\{X_{1_{i_1,\ldots,i_{k_1}}}(\boldsymbol{s},h),\ldots,X_{n^*_{i_1,\ldots,i_{k_1}}}(\boldsymbol{s},h)\right\}'\boldsymbol{C}_v^{*-1}\left\{X_{1_{i_1,\ldots,i_{k_1}}}(\boldsymbol{s},h),\ldots,X_{n^*{}_{i_1,\ldots,i_{k_1}}}(\boldsymbol{s},h)\right\}$$

Taking limit as $h\to 0$ on both sides and using (23) proves the (k_1+1)-th mean square differentiability of the process $\tilde{w}(\cdot)$ of the process at $\boldsymbol{s}$ in the direction $\boldsymbol{u}$. Since $\boldsymbol{s}$ and $\boldsymbol{u}$ are arbitrarily chosen, Predictive process is (k_1+1)-times mean square differentiable.

- For Modified Predictive Process, a simple calculation will yield

$$\mathrm{E}\left(w_{mpp}(\boldsymbol{s})-w_{mpp}(\boldsymbol{s}_0)\right)^2=2\sigma^2\{1-\boldsymbol{I}(\boldsymbol{s}=\boldsymbol{s}_0)\}\{1-\boldsymbol{c}_v(\boldsymbol{s}_1,\mathcal{S}^*)'\boldsymbol{C}_v^{*-1}\boldsymbol{c}_v(\boldsymbol{s}_1,\mathcal{S}^*)\} \tag{24}$$

thus, $\lim_{\boldsymbol{s}\to\boldsymbol{s}_0}\mathrm{E}\left(w_{mpp}(\boldsymbol{s})-w_{mpp}(\boldsymbol{s}_0)\right)^2$ does not exist. Therefore, Modified Predictive Process is not L_2 continuous.

- Let's denote $\boldsymbol{z}(\boldsymbol{s},\mathcal{S}^*)'=\boldsymbol{c}_v(\boldsymbol{s},\mathcal{S}^*)'\boldsymbol{C}_v^{*-1}$. For Tapered Predictive Process with some little algebra it can be shown that,

$$\begin{aligned}\mathrm{E}\left[w_{tap}(\boldsymbol{s})-w_{tap}(\boldsymbol{s}_0)\right]^2=2\sigma^2-\sigma^2\{&\boldsymbol{z}(\boldsymbol{s},\mathcal{S}^*)'\boldsymbol{c}_v(\boldsymbol{s},\mathcal{S}^*)(1-C_\nu(\boldsymbol{s},\boldsymbol{s}_0))\\&+C_w(\boldsymbol{s},\boldsymbol{s}_0)C_\nu(\boldsymbol{s},\boldsymbol{s}_0)\}\end{aligned} \tag{25}$$

The L_2 continuity follows easily from the equation (25).

It is well known that with matern covariance kernel with $m_1<\theta_2<m_1+1$, $w(\boldsymbol{s})$ is m_1-times differentiable anywhere. By the assumption, $C_\nu(\cdot)$ is 2k-times differentiable. Thus, Proposition 1 and (1) of Lemma 4.3 together yield $w_{tap}(\boldsymbol{s})=w_{pp}(\boldsymbol{s})+(w(\boldsymbol{s})-w_{pp}(\boldsymbol{s}))\eta(\boldsymbol{s})$ is $\min(m_1,k)$-times Mean square differentiable except at the set of knot points.

References

[1] H. Sang, M. Jun, J. Z. Huang, Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors, The Annals of Applied Statistics (2011) 2519–2548.

[2] N. Cressie, Statistics for spatial data: Wiley series in probability and statistics, Wiley-Interscience New York 15 (1993) 16.

[3] S. Banerjee, B. P. Carlin, A. E. Gelfand, Hierarchical modeling and analysis for spatial data, Crc Press, 2014.

[4] O. Schabenberger, C. A. Gotway, Statistical methods for spatial data analysis, CRC press, 2004.

[5] D. Higdon, et al., Space and space-time modeling using process convolutions, Quantitative methods for current environmental issues 3754.

[6] M. L. Stein, A modeling approach for large spatial datasets, Journal of the Korean Statistical Society 37 (1) (2008) 3–10.

[7] N. Cressie, G. Johannesson, Fixed rank kriging for very large spatial data sets, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (1) (2008) 209–226.

[8] S. Banerjee, A. E. Gelfand, A. O. Finley, H. Sang, Gaussian predictive process models for large spatial data sets, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (4) (2008) 825–848.

[9] R. Guhaniyogi, A. O. Finley, S. Banerjee, A. E. Gelfand, Adaptive gaussian predictive process models for large spatial datasets, Environmetrics 22 (8) (2011) 997–1007.

[10] M. L. Stein, Limitations on low rank approximations for covariance matrices of spatial data, Spatial Statistics 8 (2014) 1–19.

[11] A. O. Finley, S. Banerjee, P. Waldmann, T. Ericsson, Hierarchical spatial modeling of additive and dominance genetic variance for large spatial trial datasets, Biometrics 65 (2) (2009) 441–451.

[12] H. Sang, J. Z. Huang, A full scale approximation of covariance functions for large spatial data sets, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74 (1) (2012) 111–132.

[13] A. E. Gelfand, S. Banerjee, Multivariate spatial process models, Handbook of Spatial Statistics (2010) 495–515.

[14] T. Gneiting, W. Kleiber, M. Schlather, Matérn cross-covariance functions for multivariate random fields, Journal of the American Statistical Association 105 (491) (2012) 1167–1177.

[15] R. Guhaniyogi, A. O. Finley, S. Banerjee, R. K. Kobe, Modeling complex spatial dependencies: Low-rank spatially varying cross-covariances with application to soil nutrient data, Journal of Agricultural, Biological, and Environmental Statistics 18 (3) (2013) 274–298.

[16] T. V. Apanasovich, M. G. Genton, Cross-covariance functions for multivariate random fields based on latent dimensions, Biometrika 97 (1) (2010) 15–30.

[17] A. E. Gelfand, A. M. Schmidt, S. Banerjee, C. Sirmans, Nonstationary multivariate process modeling through spatially varying coregionalization, Test 13 (2) (2004) 263–312.

[18] M. G. Genton, W. Kleiber, et al., Cross-covariance functions for multivariate geostatistics, Statistical Science 30 (2) (2015) 147–163.

[19] D. Ruppert, M. P. Wand, R. J. Carroll, Semiparametric regression, no. 12, Cambridge university press, 2003.

[20] S. Banerjee, A. O. Finley, P. Waldmann, T. Ericsson, Hierarchical spatial process models for multiple traits in large genetic trials, Journal of the American Statistical Association 105 (490) (2010) 506–521.

[21] R. Furrer, M. G. Genton, D. Nychka, Covariance tapering for interpolation of large spatial datasets, Journal of Computational and Graphical Statistics.

[22] C. G. Kaufman, M. J. Schervish, D. W. Nychka, Covariance tapering for likelihood-based estimation in large spatial data sets, Journal of the American Statistical Association 103 (484) (2008) 1545–1555.

[23] T. Gneiting, Compactly supported correlation functions, Journal of Multivariate Analysis 83 (2) (2002) 493–508.

[24] J. Du, H. Zhang, V. Mandrekar, et al., Fixed-domain asymptotic properties of tapered maximum likelihood estimators, the Annals of Statistics 37 (6A) (2009) 3330–3361.

[25] D. Wang, W.-L. Loh, et al., On fixed-domain asymptotics and covariance tapering in gaussian random field models, Electronic Journal of Statistics 5 (2011) 238–269.

[26] M. Bevilacqua, A. Fassò, C. Gaetan, E. Porcu, D. Velandia, Covariance tapering for multivariate gaussian random fields estimation, Statistical Methods & Applications 25 (1) (2016) 21–37.

[27] D. J. Daley, E. Porcu, M. Bevilacqua, Classes of compactly supported covariance functions for multivariate random fields, Stochastic Environmental Research and Risk Assessment 29 (4) (2015) 1249–1263.

[28] R. J. Adler, The geometry of random fields, Vol. 62, Siam, 2010.

[29] S. Banerjee, A. Gelfand, On smoothness properties of spatial processes, Journal of Multivariate Analysis 84 (1) (2003) 85–100.

[30] K. Mardia, J. Kent, C. Goodall, J. Little, Kriging and splines with derivative information, Biometrika 83 (1) (1996) 207–221.

[31] S. Banerjee, A. E. Gelfand, Bayesian wombling: Curvilinear gradient assessment under spatial process models, Journal of the American Statistical Association 101 (476) (2006) 1487–1501.

[32] A. E. Gelfand, S. K. Ghosh, Model choice: a minimum posterior predictive loss approach, Biometrika 85 (1) (1998) 1–11.

[33] D. Mandallaz, Design-based properties of some small-area estimators in forest inventory with two-phase sampling, Canadian Journal of Forest Research 43 (5) (2013) 441–449.