Statistica Sinica **34** (2024), 837-865
doi:https://doi.org/10.5705/ss.202021.0326

TEST OF THE LATENT DIMENSION OF A SPATIAL BLIND SOURCE SEPARATION MODEL

Christoph Muehlmann, François Bachoc, Klaus Nordhausen and Mengxi Yi*

*Vienna University of Technology, Université Paul Sabatier,
University of Jyväskylä and Beijing Normal University*

Abstract: We assume a spatial blind source separation model in which the observed multivariate spatial data are a linear mixture of latent spatially uncorrelated random fields containing a number of pure white noise components. We propose a test on the number of white noise components, and obtain the asymptotic distribution of its statistic for a general domain. We also demonstrate how computations can be facilitated in the case of gridded observation locations. Based on this test, we obtain a consistent estimator of the true dimension. Simulation studies and an environmental application provided in the Supplementary Material demonstrate that our test is at least comparable to, and often outperforms bootstrap-based techniques.

Key words and phrases: Asymptotic distribution, kernel function, multivariate spatial data, signal number, spatial bootstrap.

1. Introduction

Advances in technology have led to massive amounts of multivariate spatial data being collected, for example, in the geographical, ecological (Legendre and Legendre (2012)), and atmospheric (von Storch and Zwiers (2001)) sciences. A domain expert analyzes such data by investigating and interpreting at least p maps (for the p measured variables), which might be contaminated by various sources of noise, such as measurement inconsistencies or errors. Moreover, it may be difficult to interpret the raw data because the original variable reflects a mixture of physical processes that are actually of interest. As such, we also need to investigate the dependencies between measurements. From a statisticians perspective, these spatially correlated data sets contain dependencies, both within and among the individual data processes, which makes statistical modeling of the data challenging, especially when the dimensionality p is large. With a data set of size n, it takes $cp(p+1)/2$ parameters to describe the full covariance and cross-covariance structure of the model, where c is the number of characteristic parameters per covariance and cross-covariance. Furthermore, it requires a computational cost of $\mathcal{O}(n^3p^3)$ for prediction using optimal linear predictors and for Gaussian likelihood evaluation; see Cressie (1993, Sec. 3) and

*Corresponding author.

Legendre and Legendre (2012).

One way to approach the problems arising from spatial cross-dependencies is to use the spatial blind source separation (SBSS) framework; see Nordhausen et al. (2015) and Bachoc et al. (2020b). Blind source separation (BSS) is a well-studied multivariate procedure used to recover latent variables when only a linear mixture of them is observed; see, for example, Comon and Jutten (2010) and Nordhausen and Oja (2018). A common assumption for BSS is that the latent variables are second-order stationary and uncorrelated. That is, we assume $\mathbf{x}(\mathbf{s}) = \mathbf{\Omega}\mathbf{z}(\mathbf{s})$, where $\mathbf{x}(\mathbf{s}) \in \mathbb{R}^p$ is the observed p-variate measurement at location $\mathbf{s} \in \mathbb{R}^d$, $\mathbf{z}(\mathbf{s}) \in \mathbb{R}^p$ is a latent second-order stationary p-variate source with uncorrelated components, and $\mathbf{\Omega} \in \mathbb{R}^{p\times p}$ is an unknown full-rank mixing matrix. To estimate the unmixing matrix $\mathbf{\Gamma}$, that is, $\mathbf{\Omega}^{-1}$, Nordhausen et al. (2015) propose an estimator based on the simultaneous diagonalization of two scatter matrices. Bachoc et al. (2020b) extended this method to jointly diagonalize more than two scatter matrices for multivariate spatial data. Preprocessing the data using an SBSS method is appealing from a practitioners perspective, because the latent components are more likely to reflect the physical nature of the processes that generated the data. For example, Nordhausen et al. (2015) found six meaningful physical latent components in a geostatistical data set that were not easily detectable in the original data. Moreover, it suffices to investigate only p maps, because the resulting latent components are spatially uncorrelated. Common tasks, such as modeling the spatial covariance or predictions of the original data, are again modified, because the statistical analysis can be performed using univariate tools on the latent components. The results for the latent components are transferred back to the original data because the transformation is linear in nature. Muehlmann, Nordhausen and Yi (2021) investigate this procedure in the context of geostatistical prediction. Building p univariate models rather than one multivariate model simplifies the given tasks significantly. Nevertheless, if the dimension p is still high, a further reduction may be possible if not all of the p components are of interest.

The SBSS model of Nordhausen et al. (2015) gives no preference to any of the latent components, with all p of them essentially of equal interest, from a statistical perspective. However, in practical cases of BSS, it is often assumed that only a few components are of interest and regarded as the signal, whereas the remaining components are discarded as noise. This can be represented in the statistical BSS model by supposing that the latent components consist of two parts, $\mathbf{z} = (\mathbf{z}_s^T, \mathbf{z}_w^T)^T$, where $\mathbf{z}_s \in \mathbb{R}^q$ is the signal and $\mathbf{z}_w \in \mathbb{R}^{p-q}$ is the noise. Matilainen, Nordhausen and Virta (2018), Virta and Nordhausen (2021), and Nordhausen and Virta (2018) all consider components with serial dependence as signals in a time series context. Identifying and discarding the noise part leads to fewer components needing to be investigated and modeled, which simplifies the analysis.

We consider an SBSS model in which the signals are characterized as components with second-order spatial dependence. We derive a test for the signal dimension q based on the joint diagonalization of two or more scatter matrices that are specified by kernel functions. We then provide the asymptotic distribution of the test statistic. This asymptotic result enables us to extend the framework of Bachoc et al. (2020b) to the case where the signal and noise components are not all asymptotically identifiable and their distributions are not necessarily Gaussian. We develop new proof techniques to obtain these two extensions. The first extension generalizes arguments made by Virta and Nordhausen (2021) to a spatial setting, and he second extends the arguments in Bachoc et al. (2020a) beyond the case of transformed Gaussian processes.

In addition, we demonstrate that introducing new scatter matrices compared with the one used by Bachoc et al. (2020b) enables us to obtain a neater asymptotic distribution of the test statistic (see Remark 1). Based on the test, we then provide a consistent estimator of the unknown signal dimension. Furthermore, the detection of the noise components results in a significant computational cost reduction for subsequent multivariate spatial modeling that uses only the signal components.

We propose several bootstrap versions of the test. For both the asymptotic and the bootstrap tests, we demonstrate computational gains when the observation locations are gridded. In an extensive simulation study, we show that the various tests already have levels close to the nominal level, for small to moderate sample sizes, and provide an accurate estimation of the signal dimension. We conclude that the asymptotic test is comparable to, and often outperforms the bootstrap tests, while being computationally less demanding. Employing an environmental application, we show that our methods enable the reduction of the dimension of a multivariate spatial data set, retaining the most interpretable and informative estimated independent components, and discarding the unusable ones as noise.

The remainder of the paper is organized as follows. In Section 2, we introduce the statistical setting of the problem and present our test statistic. The methods and main results are described in Section 3, and the simulation results are reported in Section 4. Section 5 concludes the paper. The proofs of the theoretical results and the environmental application are presented in the Supplementary Material.

2. Setup and Model

Suppose our data consist of a p-dimensional multivariate random field $\mathbf{x}(\mathbf{s}) = \{x_1(\mathbf{s}), \ldots, x_p(\mathbf{s})\}^T$, $\mathbf{s} \in \mathcal{S}$, where $\mathcal{S} \subseteq \mathbb{R}^d$ is a region of interest. The covariance and cross-covariance functions of $\mathbf{x}$, defining its second-order structure, are some of its central characteristics; see De Iaco et al. (2013), Genton and Kleiber (2015), and Gneiting, Kleiber and Schlather (2010) for an introduction and various approaches to modeling these functions.

Here, the second-order structure of $\mathbf{x}$ is assumed to obey an SBSS model,

$$\mathbf{x}(\mathbf{s}) = \boldsymbol{\Omega}\mathbf{z}(\mathbf{s}), \tag{2.1}$$

where $\boldsymbol{\Omega}$ is a $p \times p$ unknown invertible matrix, and $\mathbf{z}(\mathbf{s}) = \{z_1(\mathbf{s}), \ldots, z_p(\mathbf{s})\}^T$ is the latent field with independent components, with $\mathrm{Cov}(\mathbf{z}(\mathbf{s})) = \mathbf{I}_p$ for all $\mathbf{s} \in \mathcal{S}$. The SBSS model is related to a popular multivariate covariance model, namely, the linear model of coregionalization (LMC), which is expressed as

$$\mathbf{C}^{LMC}(h) = \sum_{m=1}^{r} \mathbf{T}_m \rho_m(h).$$

Here, $\mathbf{T}_m$ are nonnegative definite $p \times p$ coregionalization matrices, and $\rho_m(h)$ are univariate stationary correlation functions; see Goulard and Voltz (1992), Schmidt and Gelfand (2003), and Emery (2010). Dimension reduction in the LMC literature is performed by first fitting an LMC, and then decreasing the number of terms r or finding a lower rank representation of the coregionalization matrices. The former is achieved by using an eigendecomposition of the coregionalization matrices. If the system of eigenvectors is equal across a few summands, these matrices may be proportional. This is referred to as intrinsic correlation; see Wackernagel (1994). In the latter case, the coregionalization matrices arise from a scalar product matrix of latent variables (Goulard and Voltz (1992)). Variants of a principal component analysis (PCA) of the coregionalization matrices lead to a lower dimensional representation of these latent variables. This is called a regionalized PCA; see (Wackernagel (2003, Chap. 27)).

As noted by Bachoc et al. (2020b), the SBSS model is a special case of the LMC, where $r = p$, $\mathbf{T}_m = \boldsymbol{\omega}_m \boldsymbol{\omega}_m^\top$ ($\boldsymbol{\omega}_m$ is the mth column of the mixing matrix $\boldsymbol{\Omega}$), leading to rank-one corregionalization matrices, and $\rho_m(h)$ are the univariate correlation functions of the entries of the latent field $\mathbf{z}(\mathbf{s})$. Although there is a connection between the LMC and SBSS, the advantage of SBSS lies in the fact that estimating the unmixing matrix (or, equivalently, the coregionalization matrices) does not require estimating or specifying a model for the covariances of the latent field components. Moreover, our approach to dimension reduction is different, because we test wether some latent components are white noise, which leads to a reduction of r.

Next, we explain how to estimate the unmixing matrix $\boldsymbol{\Gamma}$, that is $\boldsymbol{\Omega}^{-1}$, and propose our test statistic for the signal dimension of the SBSS model (2.1). Let $I(\cdot)$ denote the indicator function throughout this paper, and consider the kernel functions $f_0, f_1, \ldots, f_k$, with $f_\ell : \mathbb{R}^d \to \mathbb{R}$ for $\ell = 0, \ldots, k$, and with $f_0(\mathbf{s}) = I(\mathbf{s} = \mathbf{0})$. Note that we call $f_0, f_1, \ldots, f_k$ kernels, following Bachoc et al. (2020a); Muehlmann, Bachoc and Nordhausen (2022) analogously to kernel smoothing.

However, $f_0, f_1, \ldots, f_k$ should not be confused with the covariance functions of the components of $\mathbf{x}$ or $\mathbf{z}$. For $f \in \{f_0, f_1, \ldots, f_k\}$, let

$$F_{n,f} = \frac{1}{n} \sum_{i,j=1}^{n} f^2(\mathbf{s}_i - \mathbf{s}_j),$$

where $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subseteq \mathcal{S}$ is the set of two-by-two distinct observation points. Note that $F_{n,f_0} = 1$. Let $f \in \{f_1, \ldots, f_k\}$. The population local covariance (or scatter) matrices are then defined as

$$\mathbf{M}(f) = \frac{1}{n\sqrt{F_{n,f}}} \sum_{i=1}^{n} \sum_{j=1}^{n} f(\mathbf{s}_i - \mathbf{s}_j) \mathbb{E}\left(\mathbf{x}(\mathbf{s}_i)\mathbf{x}(\mathbf{s}_j)^T\right) \tag{2.2}$$

$$\text{and} \quad \mathbf{M}(f_0) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(\mathbf{x}(\mathbf{s}_i)\mathbf{x}(\mathbf{s}_i)^T\right),$$

and the corresponding sample local covariance matrices are defined as

$$\widehat{\mathbf{M}}(f) = \frac{1}{n\sqrt{F_{n,f}}} \sum_{i=1}^{n} \sum_{j=1}^{n} f(\mathbf{s}_i - \mathbf{s}_j) \mathbf{x}(\mathbf{s}_i)\mathbf{x}(\mathbf{s}_j)^T \tag{2.3}$$

$$\text{and} \quad \widehat{\mathbf{M}}(f_0) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}(\mathbf{s}_i)\mathbf{x}(\mathbf{s}_i)^T.$$

Remark 1. The normalizing quantity $nF_{n,f}^{1/2}$ in (2.2) and (2.3) is slightly different from that in Bachoc et al. (2020b), who simply use n. Here, including $F_{n,f}^{1/2}$ enables us to obtain a simple and elegant asymptotic distribution of the test statistic for the number of noise components (see Proposition 1).

The $k+1$ sample local covariance matrices $\widehat{\mathbf{M}}(f_0), \widehat{\mathbf{M}}(f_1), \ldots, \widehat{\mathbf{M}}(f_k)$ are used to estimate the unmixing matrix $\mathbf{\Gamma}$ as

$$\widehat{\mathbf{\Gamma}} \in \underset{\substack{\mathbf{\Gamma}:\mathbf{\Gamma}\widehat{\mathbf{M}}(f_0)\mathbf{\Gamma}^T=\mathbf{I}_p, \\ \mathbf{\Gamma} \text{ has rows } \boldsymbol{\gamma}_1^T,\ldots,\boldsymbol{\gamma}_p^T, \\ (\sum_{\ell=1}^{k}\{\boldsymbol{\gamma}_j^T\widehat{\mathbf{M}}(f_\ell)\boldsymbol{\gamma}_j\}^2)_{j=1,\ldots,p} \text{ are in descending order}}}{\operatorname{argmax}} \sum_{\ell=1}^{k} \sum_{j=1}^{p} \{\boldsymbol{\gamma}_j^T \widehat{\mathbf{M}}(f_\ell) \boldsymbol{\gamma}_j\}^2. \tag{2.4}$$

The unmixing matrix should "diagonalize" all k local covariance matrices, and for $\ell = 1, \ldots, k$, we let

$$\widehat{\mathbf{D}}_\ell = \widehat{\mathbf{\Gamma}}\widehat{\mathbf{M}}(f_\ell)\widehat{\mathbf{\Gamma}}^T,$$

where all $\widehat{\mathbf{D}}_\ell$ should be close to a diagonal matrix. Note that for finite data, exact diagonalization is usually possible only for $k = 1$. Further, by definition, $\sum_{\ell=1}^{k} \widehat{\mathbf{D}}_{\ell,1,1}^2 \geq \cdots \geq \sum_{\ell=1}^{k} \widehat{\mathbf{D}}_{\ell,p,p}^2$. We are now interested in the case in which there are q "real" continuous random fields in $\mathbf{z}$, and the remaining $p - q$ components are white noise.

For $q \in \{0, \dots, p-1\}$, we are interested in testing the following hypothesis:

$$H_{0q} : \text{There are exactly } p-q \text{ white noise processes in } \mathbf{z}.$$

This hypothesis is formalized in the following two conditions:

Condition 1. *For $a = 1, \dots, p-q$, the covariance function of z_{q+a} is given by*

$$\mathrm{Cov}(z_{q+a}(\mathbf{u}), z_{q+a}(\mathbf{v})) = I(\mathbf{u} - \mathbf{v} = \mathbf{0}).$$

Condition 2. *For $\ell = 1, \dots, k, f_\ell$ is symmetric and satisfies $f_\ell(\mathbf{0}) = 0$. For $a = 1, \dots, q$, we have*

$$\liminf_{n \to \infty} \sum_{\ell=1}^{k} \left[\left(\mathbf{\Omega}^{-1} \mathbf{M}(f_\ell) \mathbf{\Omega}^{-T} \right)_{a,a} \right]^2 > 0.$$

Note that in Conditions 1 and 2, we assume that the sources are ordered such that the q signal components are first, followed by the $p-q$ noise components. Because the order of the sources is not identifiable, this assumption comes without loss of generality. The fulfillment of Condition 2 means that the correlation in the signal fields $z_1, \dots, z_q$ is sufficient for these signals to be asymptotically separated from the noise fields $z_{q+1}, \dots, z_p$. Note that we do not need to consider the stronger assumption that the q vectors of the ath diagonal elements in $\mathbf{\Omega}^{-1}\mathbf{M}(f_1)\mathbf{\Omega}^{-T}, \dots, \mathbf{\Omega}^{-1}\mathbf{M}(f_k)\mathbf{\Omega}^{-T}$, for $a = 1, \dots, q$, for the signal random fields are asymptotically distinct (see Assumption 9 in Bachoc et al. (2020b)) and nonzero.

When Condition 2 is satisfied by $f_1, \dots, f_k$, it is likely to be satisfied by the single kernel $f_1 + \cdots + f_k$ as well. This means that using a single kernel can be sufficient to obtain the various asymptotic results of Section 3 on the test statistic discussed below. Nevertheless, the flexibility of allowing several kernels is beneficial here. Indeed, after having tested (or estimated) the signal dimension, the user may be interested in estimating some of the first (most important) signal components individually. As shown in Bachoc et al. (2020b), this usually requires multiple kernels, both for theoretical guarantees and for practical efficiency. Using the same set of kernels for these two studies (signal dimension and components) may make it easier to interpret the results; see Nordhausen and Ruiz-Gazen (2022) for details on joint diagonalization in multivariate methods.

Conditions 1 and 2 motivate the following block decompositions for $\ell = 1, \dots, k$:

$$\widehat{\mathbf{M}}(f_\ell) = \begin{pmatrix} \widehat{\mathbf{M}}(f_\ell)_{qq} & \widehat{\mathbf{M}}(f_\ell)_{q0} \\ \widehat{\mathbf{M}}(f_\ell)_{0q} & \widehat{\mathbf{M}}(f_\ell)_{00} \end{pmatrix} \quad \text{and} \quad \widehat{\mathbf{D}}_\ell = \begin{pmatrix} \widehat{\mathbf{D}}_{\ell,qq} & \widehat{\mathbf{D}}_{\ell,q0} \\ \widehat{\mathbf{D}}_{\ell,0q} & \widehat{\mathbf{D}}_{\ell,00} \end{pmatrix},$$

where the blocks $\widehat{\mathbf{M}}(f_\ell)_{qq}$ and $\widehat{\mathbf{D}}_{\ell,qq}$ have size $q \times q$, and the blocks $\widehat{\mathbf{M}}(f_\ell)_{00}$ and

$\widehat{\mathbf{D}}_{\ell,00}$ have dimension $(p-q) \times (p-q)$.

Then, our test statistic is

$$t_q = \frac{n}{2} \sum_{\ell=1}^{k} ||\widehat{\mathbf{D}}_{\ell,00}||^2, \tag{2.5}$$

where $|| \cdot ||$ is the Frobenius norm. The test statistic is expected to be bounded under the null hypothesis, and to diverge when one of $z_{q+1}, \ldots, z_p$ is spatially correlated. The test will reject the null hypothesis H_{0q} if t_q is larger than a certain threshold, in which case, more than q signal components may be present. For a nominal level $\alpha \in (0,1)$, the threshold is set to the quantile $1-\alpha$ of the asymptotic distribution of Proposition 1 or 2 or Corollary 1, depending on the context.

3. Theory and Methodology

3.1. Asymptotic tests for dimension

Assume now that $\mathbf{x}$ satisfies Model (2.1). Then, let q denote the true value of the signal dimension (i.e., H_{0q} is true) and consider the limiting distribution of t_q. To establish the asymptotic results, we need to introduce a few technical conditions.

Condition 3. *The random fields $z_1, \ldots, z_p$ are independent, centered, and stationary.*

The independence assumption makes studing the sources meaningful, and the independence of the noise components is used to obtain the asymptotic distribution of the test statistic in Propositions 1 and 2 and Corollary 1; see, specifically, the proof of Proposition 1. The stationarity assumption is standard in spatial statistics; see for Shaby and Ruppert (2012) and Bachoc et al. (2020b). The zero-mean assumption is replaced by a constant unknown mean assumption in Section 3.4. For $a = 1, \ldots, p$, we let z_a have the stationary covariance function $K_a : \mathbb{R}^d \to \mathbb{R}$, with $\mathrm{Cov}(z_a(\mathbf{s}), z_a(\mathbf{s}+\mathbf{h})) = K_a(\mathbf{h})$.

Condition 4. *A fixed $\delta > 0$ exists such that, for all $n \in \mathbb{N}$ and $i \neq j, i, j = 1, \ldots, n, ||\mathbf{s}_i - \mathbf{s}_j|| \geq \delta$.*

Condition 4 implies that we have an increasing-domain asymptotic framework; see Cressie (1993, Sec. 7.3) for an introduction, and Bevilacqua et al. (2012) for recent developments.

Condition 5. *Fixed $\beta > 0$ and $\alpha > 0$ exist such that, for all $a = 1, \ldots, q$, for $u, v \in \mathbb{N}$, $u \geq 1$, $v \geq 1$, $u + v \leq 4$, for $\mathbf{y}_1, \ldots, \mathbf{y}_u \in \mathbb{R}^d$, for $\mathbf{w}_1, \ldots, \mathbf{w}_v \in \mathbb{R}^d$,*

$$|\mathrm{Cov}\left(z_a(\mathbf{y}_1) \ldots z_a(\mathbf{y}_u), z_a(\mathbf{w}_1) \ldots z_a(\mathbf{w}_v)\right)| \leq \frac{\beta}{1 + \Delta^{2d+1+\alpha}},$$

where

$$\Delta = \min_{\substack{r \in \{1,\ldots,u\}, \\ s \in \{1,\ldots,v\}}} ||\mathbf{y}_r - \mathbf{w}_s||.$$

Condition 5 means that, for the q signal processes, two products of signal values between two sets of input locations have a covariance that decays with the smallest distance between two points of the sets. Hence, this condition can be interpreted as weak dependence, and is mild in the sense that only pairs of sets with a sum of four elements or less need to be considered.

In the special case where the signal processes are stationary Gaussian, the condition holds when, for two constants $0 < \gamma_1, \gamma_2 < \infty$, and for $a = 1, \ldots, q$, $\mathbf{h} \in \mathbb{R}^d$, the covariance function satisfy $|K_a(\mathbf{h})| \leq \gamma_1 \exp(-\gamma_2 ||\mathbf{h}||)$. This can be seen from the proof of Lemma 7 in Bachoc et al. (2020a), where F is the identity function. This latter condition holds for many standard covariance functions in spatial statistics, such as the spherical, Gaussian, exponential, and Matérn functions (Cressie (1993, Sec. 2.3)). Note that the exponential decay of the covariance can be weakened to a polynomial decay, from direct arguments, and still yield Condition 5. We do not elaborate on this for the sake of brevity. Furthermore, Lemma 7 in Bachoc et al. (2020a) shows that Condition 5 holds when the signal processes are nonGaussian and obtained from nonlinear transformations of stationary Gaussian processes, under mild technical assumptions.

Note that when the signal and noise processes are stationary Gaussian, Condition 5 can be replaced by the simpler condition that, for two constants $0 < \gamma_1, \gamma_2 < \infty$, for $a = 1, \ldots, q$, $\mathbf{h} \in \mathbb{R}^d$, their covariance functions satisfy $|K_a(\mathbf{h})| \leq \gamma_1/(1 + ||\mathbf{h}||^{d+\gamma_2})$. With this replacement, Propositions 1 to 4 and Corollary 1 still hold, because in this case, Lemmas S1.3 and S1.4 in the Supplementary Material hold directly from Theorem B.1 in the supplementary material to Bachoc et al. (2020a). We skip the details for the sake of brevity.

Condition 6. *For $a = 1, \ldots, p - q$, the random variables $\{z_{q+a}(\mathbf{y}); \mathbf{y} \in \mathbb{R}^d\}$ are independent. Assuming Condition 5 holds, then for the same $\alpha > 0$, we have*

$$\max_{a=1,\ldots,p-q} \sup_{\mathbf{y} \in \mathbb{R}^d} \mathbb{E}\left(|z_{q+a}(\mathbf{y})|^{4+\alpha}\right) < \infty. \tag{3.1}$$

Condition 6 requires that the noise values be independent (not only decorrelated). The independence assumption is important for the computation of the asymptotic distribution of the test statistic, in particular, to compute moments of order more than two (see Lemma S1.6 in the Supplementary Material) and to obtain a central limit theorem (see Lemma S1.4 in the Supplementary Material). The condition in (3.1), when taken together with Condition 3 (stationarity), simply requires a finite moment of order strictly more than four for the marginal distribution of the noise, which is arguably mild.

Condition 7. *Assuming Condition* 5 *holds, then for the same* $\beta > 0$ *and* $\alpha > 0$*, we have, for* $\ell = 1, \ldots, k$,

$$|f_\ell(\mathbf{y})| \leq \frac{\beta}{1 + ||\mathbf{y}||^{d+\alpha}}.$$

A typical example of a function $f \in \{f_1, \ldots, f_k\}$ for which Conditions 2 and 7 are satisfied is the "ring" kernel:

$$R(r_1, r_2)(\mathbf{s}) = I(r_1 < ||\mathbf{s}|| \leq r_2), \tag{3.2}$$

with $0 < r_1 < r_2 < \infty$.

Condition 8. *For* $\ell = 1, \ldots, k$*, we have*

$$\liminf_{n \to \infty} F_{n,f_\ell} > 0.$$

Condition 8 is mild and simply requires that for $\ell = 1, \ldots, k$, the number of pairs of observation locations $\mathbf{s}_i, \mathbf{s}_j, i, j = 1, \ldots, n$, for which $f_\ell(\mathbf{s}_i - \mathbf{s}_j)$ is nonzero is not negligible compared with n.

Condition 9. *For all* $\ell, \ell^{'} = 1, \ldots, k, \ell \neq \ell^{'}, f_\ell(\mathbf{y}) f_{\ell'}(\mathbf{y}) = 0$*, for all* $\mathbf{y} \in \mathbb{R}^d$.

Condition 9 means that the supports of the kernels are disjoint. This enables us to have a simple and elegant chi-squared asymptotic distribution of the test statistic. When Condition 9 does not hold, we can still compute the asymptotic distribution of the test statistic (see Proposition 2), which is less simple, but still explicit. Hence, importantly, Condition 9 is not necessary to have an asymptotically valid test when the quantiles from the asymptotic null distribution are simple to approximate numerically. As discussed above, the kernels in Condition 9 are not the covariance functions of $\mathbf{x}$ or $\mathbf{z}$, so Condition 9 does not make any assumption on the covariance structures of $\mathbf{x}$ and $\mathbf{z}$.

Our first main result is on the asymptotic null distribution of our test statistic t_q.

Proposition 1. *Assume that Conditions* 1–9 *hold. Then, as* $n \to \infty$,

$$t_q \xrightarrow{d} \chi^2_{k(p-q)(p-q+1)/2}.$$

In the next proposition, we show that when considering the same normalization as that considered by Bachoc et al. (2020b) for the local covariance matrices, and removing the assumption of disjoint kernel supports, we still obtain an asymptotic distribution of the test statistic as the distribution of the squared Euclidean norm of a Gaussian vector. In this proposition, we consider a metric d_w generating the topology of weak convergence on the set of Borel probability measures on Euclidean spaces (e.g., Dudley (2018, p.393)).

Proposition 2. *Assume that Conditions* 1–7 *hold. Let the test statistic* $\tilde{t}_q$ *be defined as* t_q, *with* $\widehat{\mathbf{M}}(f)$ *replaced by*

$$\widetilde{\mathbf{M}}(f) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} f(\mathbf{s}_i - \mathbf{s}_j)\mathbf{x}(\mathbf{s}_i)\mathbf{x}(\mathbf{s}_j)^T,$$

for $f \in \{f_1, \ldots, f_k\}$. *Let* $\mathcal{L}_{\tilde{t}_q,n}$ *be the distribution of the test statistic* $\tilde{t}_q$, *and let* $\mathcal{L}_{\mathbf{V},n}$ *be the distribution of* $\sum_{\ell=1}^{k}\sum_{a,b=1}^{p-q}\mathbf{V}_{\ell,a,b}^2$, *where* $(\mathbf{V}_{\ell,a,b})_{\ell=1,\ldots,k;a,b=1,\ldots,p-q}$ *is a Gaussian vector with mean vector* $\mathbf{0}$ *and covariance matrix defined by*

$$\mathrm{Cov}(\mathbf{V}_{\ell,a,b}, \mathbf{V}_{\ell',a',b'}) = \frac{1}{2}F_{n,f_\ell,f_{\ell'}}(I(a=a')I(b=b') + I(a=b')I(b=a')),$$

with

$$F_{n,f_\ell,f_{\ell'}} = \frac{1}{n}\sum_{i,j=1}^{n} f_\ell(\mathbf{s}_i - \mathbf{s}_j)f_{\ell'}(\mathbf{s}_i - \mathbf{s}_j),$$

for $\ell, \ell' = 1, \ldots, k$ *and* $a, b, a', b' = 1, \ldots, p-q$. *Then, as* $n \to \infty$,

$$d_w(\mathcal{L}_{\tilde{t}_q,n}, \mathcal{L}_{\mathbf{V},n}) \to 0.$$

In the following corollary, we show that if the supports of the kernels are disjoint, the test statistic converges to a weighted chi-squared distribution; see, for example, Bodenham and Adams (2016) for the approximation procedures for this distribution.

Corollary 1. *Consider the setting of Proposition* 2 *and assume additionally that Condition* 9 *holds. Then, the limiting distribution* $\mathcal{L}_{\mathbf{V},n}$ *in Proposition* 2 *is equal to the distribution of*

$$\sum_{\ell=1}^{k} F_{n,f_\ell}\mathcal{X}_\ell^2,$$

where $\mathcal{X}_1^2, \ldots, \mathcal{X}_k^2$ *are independent and follow a chi-squared distribution with* $(p-q)(p-q+1)/2$ *degrees of freedom.*

3.2. Regular domain as a special example

When the data are observed in a regular-grid domain, that is, $\mathcal{S} \subseteq \mathbb{Z}^d$, the kernel functions can be based on the natural notion of a spatial neighborhood on the grid, which simplifies our technique.

A location $\mathbf{s}_0 = (s_1, \ldots, s_d) \in \mathbb{Z}^d$ has $2d$ one-way lag-h neighbors, $(s_1 \pm h, \ldots, s_d), (s_1, s_2 \pm h, \ldots, s_d), \ldots, (s_1, \ldots, s_{d-1}, s_d \pm h)$. For example, if $d = 2$ and $h = 1$, the four one-way lag-1 neighbors of $\mathbf{s}_0$ are “left” $(s_1 - 1, s_2)$, “right” $(s_1 + 1, s_2)$, “up” $(s_1, s_2 + 1)$, and “down” $(s_1, s_2 - 1)$. Therefore, we can define the one-way lag-1 population and sample local covariance matrices as

$$\mathbf{M} = \frac{1}{\sqrt{n \sum_{i=1}^{n} |\mathcal{N}_{\mathbf{s}_i}|}} \sum_{i=1}^{n} \sum_{\mathbf{s}_j \in \mathcal{N}_{\mathbf{s}_i}} \mathbb{E}\left(\mathbf{x}(\mathbf{s}_i)\mathbf{x}(\mathbf{s}_j)^T\right)$$
$$\text{and } \widehat{\mathbf{M}} = \frac{1}{\sqrt{n \sum_{i=1}^{n} |\mathcal{N}_{\mathbf{s}_i}|}} \sum_{i=1}^{n} \sum_{\mathbf{s}_j \in \mathcal{N}_{\mathbf{s}_i}} \mathbf{x}(\mathbf{s}_i)\mathbf{x}(\mathbf{s}_j)^T, \tag{3.3}$$

respectively, where, for $\mathbf{x} \in \mathbb{Z}^d$,

$$\mathcal{N}_{\mathbf{x}} = \{\mathbf{s} \in \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}; |\mathbf{x} - \mathbf{s}| = 1\},$$

with $|\boldsymbol{u}| = |u_1| + \cdots + |u_d|$ for $\boldsymbol{u} = (u_1, \ldots, u_d) \in \mathbb{R}^d$. The matrices $\mathbf{M}$ and $\widehat{\mathbf{M}}$ are of the form $\mathbf{M}(f)$ and $\widehat{\mathbf{M}}(f)$ in (2.2) and (2.3) for $f(\mathbf{s}) = I(||\mathbf{s}|| = 1)$, $\mathbf{s} \in \mathbb{R}^d$.

Similarly, if $d = 2$, a location $\mathbf{s}_0 = (s_1, s_2) \in \mathbb{Z}^2$ has four two-way lag-1 neighbors that are of the form $(s_1 \pm 1, s_2 \pm 1)$. In general, for $m, h \in \mathbb{N}$, $1 \le m \le d$, the m-way lag-h population and sample local covariance matrices can be defined as

$$\mathbf{M} = \frac{1}{\sqrt{n \sum_{i=1}^{n} |\mathcal{N}_{h,\mathbf{s}_i}^m|}} \sum_{i=1}^{n} \sum_{\mathbf{s}_j \in \mathcal{N}_{h,\mathbf{s}_i}^m} \mathbb{E}\left(\mathbf{x}(\mathbf{s}_i)\mathbf{x}(\mathbf{s}_j)^T\right)$$
$$\text{and } \widehat{\mathbf{M}} = \frac{1}{\sqrt{n \sum_{i=1}^{n} |\mathcal{N}_{h,\mathbf{s}_i}^m|}} \sum_{i=1}^{n} \sum_{\mathbf{s}_j \in \mathcal{N}_{h,\mathbf{s}_i}^m} \mathbf{x}(\mathbf{s}_i)\mathbf{x}(\mathbf{s}_j)^T, \tag{3.4}$$

respectively, where, for $\mathbf{x} \in \mathbb{Z}^d$,

$$\mathcal{N}_{h,\mathbf{x}}^m = \{\mathbf{s} \in \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}; \mathbf{s} = \psi_J(\mathbf{x}, \zeta_J(\mathbf{x}) + h\boldsymbol{v}), \text{ for some } J \in \mathcal{A}_m, \boldsymbol{v} \in \{-1, 1\}^m\},$$

with $\mathcal{A}_m = \{J = (i_1, \ldots, i_m) \in \mathbb{N}^m; 1 \le i_1 < \cdots < i_m \le d\}$; that is, $|\mathcal{A}_m| = \binom{d}{m}$ and, for $J = (i_1, \ldots, i_m) \in \mathcal{A}_m, \mathbf{y} = (y_1, \ldots, y_m) \in \mathbb{Z}^m, \zeta_J(\mathbf{x}) = (x_{i_1}, \ldots, x_{i_m})$, $\psi_J(\mathbf{x}, \mathbf{y}) = (x_1, \ldots, x_{i_1 - 1}, y_1, x_{i_1 + 1}, \ldots, x_{i_m - 1}, y_m, x_{i_m + 1}, \ldots, x_d)$.

In general, in Equations (2.2) and (2.3), the m-way lag-h population and sample local covariance matrices $\mathbf{M}$ and $\widehat{\mathbf{M}}$, respectively, can also be written in the form $\mathbf{M}(f)$ and $\widehat{\mathbf{M}}(f)$, respectively, with $f(\mathbf{s}) = I(\mathbf{s} \in \{-h, 0, h\}^d, |\mathbf{s}| = hm)$, $\mathbf{s} \in \mathbb{R}^d$.

Consequently, for the similarly defined test statistic t_q, we can derive the same limiting conclusions. By exploiting the neighborhood structure in the regular domain case, we can also shorten the computation time for other techniques, such as the proposed asymptotic test or spatial bootstrap, with the greatest time improvement being achieved for the latter (see Sections 3.5 and 4.3).

3.3. Estimation of the number of signal components

In this section, we investigate an estimator of the signal number q based on the asymptotic tests. We wish to test the null hypothesis, for $r \in \{0, \ldots, p-1\}$,

$$H_{0r} : \text{There are exactly } p-r \text{ white noise processes in } \mathbf{z}.$$

This hypothesis states that the signal dimension is r. Similarly to Section 2, for $r = 0, \ldots, p-1$, we can partition, for $\ell = 1, \ldots, k$,

$$\widehat{\mathbf{D}}_\ell = \begin{pmatrix} \widehat{\mathbf{D}}_{\ell,rr} & \widehat{\mathbf{D}}_{\ell,r-r} \\ \widehat{\mathbf{D}}_{\ell,-rr} & \widehat{\mathbf{D}}_{\ell,-r-r} \end{pmatrix},$$

where the block $\widehat{\mathbf{D}}_{\ell,rr}$ has size $r \times r$ and the block $\widehat{\mathbf{D}}_{\ell,-r-r}$ has size $(p-r) \times (p-r)$. Then, consider the test statistic

$$t_r = \frac{n}{2} \sum_{\ell=1}^{k} ||\widehat{\mathbf{D}}_{\ell,-r-r}||^2.$$

We now use the test statistic t_r, for $r = 0, 1, \ldots, p-1$, for the estimation problem and derive a number of useful limiting properties in the following proposition.

Proposition 3. *Assume the same conditions as in Proposition* 1. *Then,*

- *If $r \geq q$, then t_r is bounded in probability.*
- *If $r < q$, then there exists a fixed $b > 0$ such that $t_r/n \geq b + o_p(1)$.*

A consistent estimate $\hat{q}$ of the unknown signal dimension $q \leq p-1$ can then be based on the test statistic t_r, as the following proposition states.

Proposition 4. *Assume the same conditions as in Proposition* 1. *Let $(c_n)_{n \in \mathbb{N}}$ be a sequence of positive numbers such that $c_n \to \infty$ and $c_n = o(n)$ as $n \to \infty$. Let*

$$\hat{q} = \min \{ r \in \{1, \ldots, p-1\} \,|\, t_r \leq c_n \},$$

with the convention $\min \varnothing = p$. *Then, $\hat{q} \to q$ in probability as $n \to \infty$.*

However, specifying the sequence c_n is not obvious in practice, and is still an open problem in order determination using hypothesis tests based on eigenvalues, as can be done a PCA, sliced inverse regression, or independent components analysis; see, for example, Bura and Cook (2001), Nordhausen et al. (2017) and Nordhausen, Oja and Tyler (2022), and the reference therein. Nevertheless, an estimator $\hat{q}$ can also be found by applying a suitable strategy to perform successive tests. Later, in the simulations, we always test for simplicity at the same significance level, and apply a divide-and-conquer strategy for the testing.

Remark 2. One can check that Propositions 3 and 4 still hold if the setting of Proposition 1 is replaced by that of Proposition 2.

3.4. General mean

The previous results were derived under the assumption that $\mathbb{E}(\mathbf{z}(\mathbf{s})) = \mathbf{0}$. In the next proposition, we show that the conclusions of Propositions 1, 2, 3, and 4 and Corollary 1 are unchanged when $\mathbf{z}$ has a nonzero unknown constant mean function, and when the observations are empirically centered for the computation of the local covariance matrices.

Proposition 5. *Assume that for* $a = 1, \dots, p, z_a$ *has a constant mean function* $\mu_a \in \mathbb{R}$*. For* $f \in \{f_1, \dots, f_k\}$*, let*

$$\overline{\mathbf{M}}(f) = \frac{1}{n\sqrt{F_{n,f}}} \sum_{i=1}^{n} \sum_{j=1}^{n} f(\mathbf{s}_i - \mathbf{s}_j)(\mathbf{x}(\mathbf{s}_i) - \bar{\mathbf{x}})(\mathbf{x}(\mathbf{s}_j) - \bar{\mathbf{x}})^T$$

$$\text{and} \quad \mathbf{M}(f_0) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}(\mathbf{s}_i) - \bar{\mathbf{x}})(\mathbf{x}(\mathbf{s}_i) - \bar{\mathbf{x}})^T, \tag{3.5}$$

with $\bar{\mathbf{x}} = (1/n) \sum_{i=1}^{n} \mathbf{x}(\mathbf{s}_i)$*. Then, the conclusions of Propositions* 1, 2, 3, *and* 4 *and Corollary* 1 *still hold under the same assumptions, except that* $\widehat{\mathbf{M}}(f)$ *is replaced everywhere with* $\overline{\mathbf{M}}(f)$*.*

For the remainder of the paper, we assume that the mean is unknown.

3.5. Bootstrap tests for dimension

The above derived noise dimension test based on the large-sample behavior of the introduced test statistic is efficient to compute, but may need a large sample size for the finite-sample level to match the asymptotic level. As an alternative for smaller sample sizes, we can formulate noise dimension tests using the bootstrap method.

In its original form, the bootstrap is a nonparametric tool for estimating the distribution of an estimator or test statistic by resampling from the empirical cumulative distribution function (ECDF) of the sample at hand. It exhibits good performance in many theoretical and practical statistical problems; see Chernick et al. (2011) or Lahiri (2003) for a more detailed discussion.

Again, we assume that the observed random field follows the SBSS model given by Equation (2.1), and want to test H_{0r} given an SBSS solution of Equation (2.4) for a certain kernel setting and the corresponding test statistic in Equation (2.5). In the following, we formulate a method for resampling from the distribution of Model (2.1) by respecting the null hypothesis H_{0r}. FollowingMatilainen, Nordhausen and Virta (2018), this is achieved by leaving the hypothetical signal part of the estimated latent field $\hat{\mathbf{z}}(\mathbf{s}) = \hat{\mathbf{\Gamma}}\mathbf{x}(\mathbf{s})$ untouched, and then manipulating only the hypothetical noise parts $(\hat{\mathbf{z}}(\mathbf{s}))_i$, for $i = r + 1, \dots, p$ and all $\mathbf{s} \in \{\mathbf{s}_1, \dots, \mathbf{s}_n\}$, in one of the following ways.

Parametric: Here, we assume that each noise part is independent and identically distributed (i.i.d.) Gaussian, as is usual for white noise processes. This leads to bootstrap samples $(\mathbf{z}^*(\mathbf{s}))_i \sim N(0,1)$, for $i = r+1, \ldots, p$ and corresponding to each $\mathbf{s} \in \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$.

Permute: Here, we assume that each noise component is still i.i.d. but that it does not necessarily follow a Gaussian distribution. Therefore, bootstrap samples are drawn from the ECDF of the joint noise components: $(\mathbf{z}^*(\mathbf{s}))_i \sim \text{ECDF}((\hat{\mathbf{z}}(\mathbf{s}_1)^\top)_{\hat{w}}, \ldots, (\hat{\mathbf{z}}(\mathbf{s}_n)^\top)_{\hat{w}})$, with $i = r+1, \ldots, p$, $\mathbf{s} \in \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, and where $\hat{w}$ denotes the noise components ($r+1$ to p) of $\hat{\mathbf{z}}$.

Algorithm 1 Testing $H_{0r} : q = r$.

```
Set the number of resamples $B$, the observed sample $\mathbf{X} = (\mathbf{x}(\mathbf{s}_1), \ldots, \mathbf{x}(\mathbf{s}_n))^\top$, the flag *spatial_resampling*, and optionally, the block size $m$;
Compute the SBSS solution and get $\hat{\mathbf{\Gamma}}$ and $\hat{\mathbf{Z}} = (\hat{\mathbf{\Gamma}}\mathbf{X}^\top)^\top$ and compute the test statistic $t = t_r(\mathbf{X})$;
for $k \in \{1, \ldots, B\}$ do
    Replace the last $p-r$ columns of $\hat{\mathbf{Z}}$ with either a parametric or bootstrap sample to get $\mathbf{Z}^{*k}$;
    if *spatial_resampling* = *TRUE* then
        Replace $\mathbf{Z}^{*k}$ by a full spatial bootstrap sample. See text for details.
    Compute $\mathbf{X}^{*k} \leftarrow \hat{\mathbf{\Gamma}}\mathbf{Z}^{*k}$ and $t^k \leftarrow t_r(\mathbf{X}^{*k})$;
Return the $p$-value: $[\#(t^k \geq t) + 1]/(B+1)$;
```

After replacing the hypothetical noise part with a bootstrap sample in one of the former ways, we achieve the goal of sampling from Model (2.1) under H_{0r}. However, so far, the uncertainty of estimating the signal has not been considered in the bootstrap test. Therefore, an optional second step in the resampling procedure is devoted to drawing a spatial bootstrap sample from the already manipulated sample, as follows. We suggest applying spatial bootstrapping, as discussed in Lahiri (2003), and in the following, we summarize the main ideas. Recall that the set of sampling sites $\mathcal{C} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ lies inside the d-dimensional spatial domain $\mathcal{S}$, which can be viewed as the "sample region", and hence $\mathcal{C} \subseteq \mathcal{S} \subseteq \mathbb{R}^d$. $\mathcal{S}$ is divided into nonoverlapping blocks of size m^d that lie partially in $\mathcal{S}$, formally $\mathcal{B} = \{b_{\mathbf{i}} = (\mathbf{i}+(0,1]^d)m \cap \mathcal{S} : (\mathbf{i}+(0,1]^d)m \cap \mathcal{S} \neq \emptyset, \mathbf{i} \in \mathbb{Z}^d\}$, and overlapping blocks that lie fully in $\mathcal{S}$, written as $\mathcal{B}_{bs} = \{b_{\mathbf{j}} = \mathbf{j} + (0,1]^d m : \mathbf{j} + (0,1]^d m \subseteq \mathcal{S}, \mathbf{j} \in \mathbb{Z}^d\}$. The bootstrapped spatial domain $\mathcal{S}^*$ is formed by replacing each block $b_{\mathbf{i}} \in \mathcal{B}$ with a randomly with-replacement sampled block $b_{\mathbf{j}} \in \mathcal{B}_{bs}$ that is trimmed to the shape of $b_{\mathbf{i}}$ by $b_{\mathbf{j}} \cap (b_{\mathbf{i}} - \mathbf{i}m + \mathbf{j})$. Hence, the trimmed version of $b_{\mathbf{j}}$ remains at the original location of $b_{\mathbf{j}}$, and the shape changes to that of $b_{\mathbf{i}}$, taking care of the boundary blocks that do not lie fully within $\mathcal{S}$. Finally, the bootstrapped version of the random field is expressed as $\mathbf{z}^* = \{\mathbf{z}(\mathbf{s}) : \mathbf{s} \in \mathcal{S}^* \cap \mathcal{C}\}$. Note that in each spatial bootstrap iteration, the shape of $\mathcal{S}^*$, and therefore the bootstrapped sampling sites, differ. This, in turn, makes the computation of the local covariance matrices a demanding task, because it relies on the distances between all sampling sites, which need to be computed in each iteration. For regular data, this can be avoided by using a slightly different bootstrap regime, as follows.

Algorithm 2 Divide and Conquer.

```
Set $lower$, $upper$, and $\alpha$;
$middle = \lfloor (upper - lower)/2 \rfloor$;
while ($middle != lower$) && ($middle != upper$) do
    $p = test\_function(r = middle)$;
    if $p < alpha$ then
        $lower = middle$;
    else
        $upper = middle$;
    $middle = \lfloor (upper - lower)/2 \rfloor$;
Return $\hat{q} = middle + 1$;
```

Nordman, Lahiri and Fridley (2007) have suggested a slightly different approach for sampling sites located on a regular grid, meaning that the sampling sites satisfy $\{\mathbf{s}_1, \dots, \mathbf{s}_n\} \subseteq \mathcal{S} \cap \mathbb{Z}^d$. Again, the domain $\mathcal{S}$ is divided into blocks of size m^d that are either nonoverlapping or overlapping, but lie completely inside $\mathcal{S}$, leading to $\mathcal{B} = \{(\mathbf{i} + (0,1]^d)m : (\mathbf{i} + (0,1]^d)m \subseteq \mathcal{S}, \mathbf{i} \in \mathbb{Z}^d\}$ and $\mathcal{B}_{bs}$, as defined above. The key difference is that the bootstrap sample is drawn at the level of the random field values, whereas the former bootstrap version operates at the level of the spatial domain. Specifically, for each block $b_{\mathbf{i}} \in \mathcal{B}$, the values $\{\mathbf{z}(\mathbf{s}) : \mathbf{s} \in b_{\mathbf{i}} \cap \mathbb{Z}^d\}$ are replaced with $\{\mathbf{z}(\mathbf{s}) : \mathbf{s} \in b_{\mathbf{j}} \cap \mathbb{Z}^d\}$ for a randomly with-replacement chosen block $b_{\mathbf{j}} \in \mathcal{B}_{bs}$. This procedure keeps the bootstrapped spatial domain and sampling sites equal in all iterations, namely, the unison of all blocks from $\mathcal{B}$. This, in turn, simplifies the computation of the local covariance matrices, because only the random field values change. We compare the computation times of the former two approaches in the simulation study presented in Section 4.3.

Algorithm 1 summarizes the bootstrap strategy to test for one specific value of signal dimension r. To estimate the signal dimension, we perform a sequence of tests for different signal dimensions r at a given significance level α. Several different test sequences are possible, but we rely on the divide-and-conquer strategy outlined in Algorithm 2. Here, the $test\_function$ can be one of the bootstrap test variants seen in Algorithm 1 or the asymptotic test outlined above.

4. Simulation

To validate the performance of the proposed methods, we carried out five extensive simulation studies in `R 3.6.1` R Core Team (2019) with the help of the packages `SpatialBSS` from Muehlmann, Nordhausen and Virta (2020), `JADE` from Miettinen, Nordhausen and Taskinen (2017), `sp` from Bivand, Pebesma and Gomez-Rubio (2013), `raster` from Hijmans (2020), `gstat` from Gräler, Pebesma and Heuvelink (2016), and `RandomFields` from Schlather et al. (2015).

4.1. Simulation study 1: Hypothesis testing

In this part of the simulation, we explore the performance of the hypothesis testing. For all the following simulations, we consider the SBSS model in Equation (2.1), where without loss of generality, we set $\mu_a = 0$, for $a = 1, \ldots, p$, and assume the mean is unknown. For the latent signal part, we use two different three-variate random field model settings. Therefore, the true dimension is always $q = 3$. All the random fields are Gaussian and follow a Matérn correlation structure; thus, the ath random field z_a has its covariance function value at $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, given by

$$K_a(h; \nu, \phi) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{h}{\phi}\right)^{\nu} K_\nu \left(\frac{h}{\phi}\right), \quad h = ||\boldsymbol{u} - \boldsymbol{v}||,$$

where $\nu > 0$ is the shape parameter, $\phi > 0$ is the range parameter, K_ν is the modified Bessel function of the second kind with shape parameter ν, and Γ is the gamma function. The parameters used are $(\nu, \phi) \in \{(3, 2), (2, 1.5), (1, 1)\}$ and $\{(3, 2), (2, 1.5), (0.6, 0.6)\}$ for model settings 1 and 2, respectively, which are depicted in Figure 1. Model setting 2 can be viewed as a low-dependence version of model setting 1. The noise part always consists of i.i.d. samples drawn from $N_2(\mathbf{0}, \mathbf{I}_2)$, leading to a total latent field dimension of $p = 5$ for both model settings. Because SBSS is affine equivariant (for details, see Bachoc et al. (2020b) and the Supplementary Material), we choose the mixing matrix to be the identity matrix, that is, $\boldsymbol{\Omega} = \mathbf{I_5}$, without loss of generality.

We focus on squared spatial domains $[0, n] \times [0, n]$ (also written in the following as $n \times n$) of different sizes $n \in \{30, 40, 50, 60\}$. For a given domain, we consider two sample location patterns: uniform and skewed. For the uniform pattern, n^2 pairs of (x, y)-coordinates are drawn randomly from a uniform distribution $U(0, 1)$ and then multiplied by n, leading to a constant sampling location density over the entire domain. We follow the same approach for the skewed pattern, except that the x-coordinate values are drawn from a beta distribution $\beta(2, 5)$, resulting in a denser arrangement of samples in the left half of the domain.

For the local covariance matrices (2.3), we use two kernel function settings. Kernel setting 1 uses only one ring kernel function (3.2) with parameters $(r_1, r_2) = (0, 2)$, and kernel setting 2 uses three ring kernel functions with parameters $(r_1, r_2) \in \{(0, 2), (2, 4), (4, 6)\}$. Figure 1 depicts a simulation example for each of the uniform and skewed coordinate patterns, where the circles represent the different ring kernel radii.

Table 1. Rejection rates for model setting 1 based on 2,000 simulation repetitions at a significance level of $\alpha = 0.05$.

		Uniform						Skew					
		Kernel Setting 1			Kernel Setting 2			Kernel Setting 1			Kernel Setting 2		
Domain	Method	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}
30×30	Asym	1.000	0.041	0.006	1.000	0.042	0.007	1.000	0.042	0.004	1.000	0.029	0.003
	Sp Param	1.000	0.048	0.006	1.000	0.058	0.001	1.000	0.059	0.004	1.000	0.051	0.000
	Sp Perm	1.000	0.050	0.006	1.000	0.059	0.000	1.000	0.058	0.004	1.000	0.052	0.001
	Param	1.000	0.042	0.006	1.000	0.044	0.006	1.000	0.050	0.008	1.000	0.039	0.005
	Perm	1.000	0.045	0.008	1.000	0.051	0.006	1.000	0.049	0.008	1.000	0.035	0.005
40×40	Asym	1.000	0.055	0.004	1.000	0.048	0.005	1.000	0.045	0.002	1.000	0.040	0.005
	Sp Param	1.000	0.056	0.005	1.000	0.066	0.000	1.000	0.056	0.003	1.000	0.064	0.002
	Sp Perm	1.000	0.063	0.005	1.000	0.061	0.000	1.000	0.055	0.004	1.000	0.065	0.002
	Param	1.000	0.052	0.007	1.000	0.055	0.003	1.000	0.050	0.007	1.000	0.048	0.005
	Perm	1.000	0.056	0.007	1.000	0.052	0.004	1.000	0.048	0.008	1.000	0.050	0.004
50×50	Asym	1.000	0.049	0.005	1.000	0.040	0.010	1.000	0.040	0.006	1.000	0.044	0.009
	Sp Param	1.000	0.052	0.004	1.000	0.053	0.002	1.000	0.047	0.006	1.000	0.064	0.002
	Sp Perm	1.000	0.050	0.005	1.000	0.053	0.002	1.000	0.045	0.005	1.000	0.061	0.002
	Param	1.000	0.052	0.007	1.000	0.049	0.007	1.000	0.042	0.007	1.000	0.054	0.007
	Perm	1.000	0.050	0.008	1.000	0.050	0.006	1.000	0.042	0.010	1.000	0.054	0.008
60×60	Asym	1.000	0.052	0.006	1.000	0.048	0.010	1.000	0.044	0.004	1.000	0.045	0.004
	Sp Param	1.000	0.056	0.006	1.000	0.058	0.003	1.000	0.048	0.005	1.000	0.060	0.000
	Sp Perm	1.000	0.055	0.007	1.000	0.057	0.002	1.000	0.052	0.004	1.000	0.058	0.000
	Param	1.000	0.049	0.009	1.000	0.054	0.006	1.000	0.043	0.006	1.000	0.048	0.004
	Perm	1.000	0.053	0.009	1.000	0.050	0.008	1.000	0.046	0.006	1.000	0.048	0.004

Table 2. Rejection rates for model setting 2 based on 2,000 simulation repetitions at a significance level of $\alpha = 0.05$.

		Uniform						Skew					
		Kernel Setting 1			Kernel Setting 2			Kernel Setting 1			Kernel Setting 2		
Domain	Method	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}
30×30	Asym	1.000	0.051	0.005	1.000	0.052	0.004	1.000	0.048	0.006	1.000	0.033	0.003
	Sp Param	1.000	0.053	0.005	1.000	0.062	0.000	1.000	0.058	0.005	1.000	0.055	0.002
	Sp Perm	1.000	0.052	0.006	1.000	0.065	0.001	1.000	0.056	0.006	1.000	0.051	0.001
	Param	1.000	0.052	0.011	1.000	0.058	0.003	1.000	0.059	0.011	1.000	0.043	0.004
	Perm	1.000	0.048	0.011	1.000	0.060	0.002	1.000	0.061	0.012	1.000	0.044	0.003
40×40	Asym	1.000	0.060	0.004	1.000	0.052	0.005	1.000	0.050	0.004	1.000	0.038	0.007
	Sp Param	1.000	0.063	0.002	1.000	0.060	0.000	1.000	0.060	0.004	1.000	0.054	0.002
	Sp Perm	1.000	0.055	0.002	1.000	0.062	0.000	1.000	0.058	0.002	1.000	0.057	0.002
	Param	1.000	0.056	0.006	1.000	0.056	0.004	1.000	0.052	0.008	1.000	0.045	0.005
	Perm	1.000	0.058	0.005	1.000	0.053	0.004	1.000	0.054	0.006	1.000	0.045	0.005
50×50	Asym	1.000	0.045	0.004	1.000	0.047	0.004	1.000	0.044	0.005	1.000	0.044	0.004
	Sp Param	1.000	0.048	0.002	1.000	0.056	0.000	1.000	0.053	0.002	1.000	0.058	0.001
	Sp Perm	1.000	0.049	0.002	1.000	0.053	0.001	1.000	0.050	0.005	1.000	0.055	0.001
	Param	1.000	0.045	0.004	1.000	0.050	0.002	1.000	0.048	0.007	1.000	0.051	0.004
	Perm	1.000	0.044	0.007	1.000	0.048	0.003	1.000	0.046	0.009	1.000	0.052	0.004
60×60	Asym	1.000	0.048	0.004	1.000	0.059	0.008	1.000	0.047	0.004	1.000	0.042	0.006
	Sp Param	1.000	0.052	0.005	1.000	0.072	0.002	1.000	0.050	0.004	1.000	0.059	0.000
	Sp Perm	1.000	0.056	0.003	1.000	0.068	0.002	1.000	0.050	0.004	1.000	0.057	0.000
	Param	1.000	0.047	0.009	1.000	0.063	0.004	1.000	0.046	0.005	1.000	0.052	0.003
	Perm	1.000	0.048	0.010	1.000	0.063	0.006	1.000	0.048	0.005	1.000	0.050	0.005

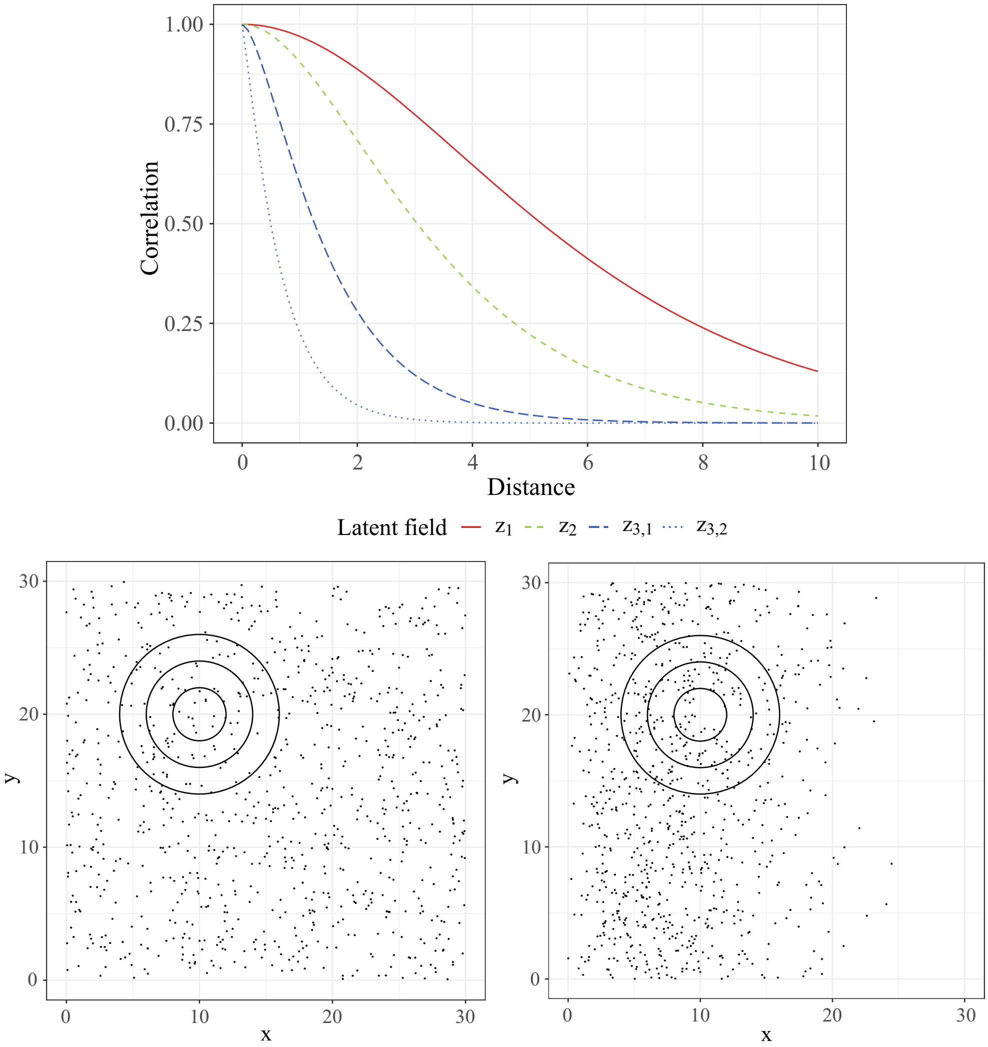

Figure 1. Upper: Matérn correlation functions for model setting 1, which consists of the signal random field $(z_1, z_2, z_{3,1})$ with parameters $(\nu, \phi) \in \{(3,2),(2,1.5),(1,1)\}$, and model setting 2 formed by the signal random field $(z_1, z_2, z_{3,2})$ with parameters $(\nu, \phi) \in \{(3,2),(2,1.5),(0.6,0.6)\}$. Lower left and lower right: uniform (left) and skewed (right) coordinate sample pattern for a spatial domain of size 30×30 with three circles of radii $(2,4,6)$ representing ring kernel functions.

For each of the four simulation settings, we perform 2,000 repetitions, and in each repetition, we test three null hypotheses (H_{02}, H_{03}, and H_{04}) using the following five test approaches: asymptotic test (Asym); noise bootstrapping with option parametric (Param); noise bootstrapping with option permute (Perm); full spatial bootstrapping with option parametric (Sp Param); and full spatial bootstrapping with option permute (Sp Perm). For all bootstrap approaches, we fix the number of resamples to $B = 200$, and for the full spatial bootstrap, the

Table 3. Rejection rates for model setting 1 with Gaussian and nonGaussian distributions and the uniform sample location pattern based on 2,000 simulation repetitions at a significance level of $\alpha = 0.05$.

		Kernel Setting 1						Kernel Setting 2					
		Gaussian			Non-Gaussian			Gaussian			Non-Gaussian		
Domain	Method	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}
30×30	Asym	1.000	0.047	0.007	1.000	0.044	0.006	1.000	0.045	0.005	1.000	0.043	0.005
	Sp Perm	1.000	0.050	0.006	1.000	0.056	0.005	1.000	0.068	0.001	1.000	0.059	0.002
	Perm	1.000	0.050	0.009	1.000	0.040	0.008	1.000	0.058	0.005	1.000	0.048	0.005
40×40	Asym	1.000	0.038	0.002	1.000	0.040	0.002	1.000	0.048	0.006	1.000	0.042	0.008
	Sp Perm	1.000	0.043	0.002	1.000	0.044	0.002	1.000	0.056	0.000	1.000	0.056	0.002
	Perm	1.000	0.038	0.006	1.000	0.046	0.007	1.000	0.049	0.003	1.000	0.046	0.005
50×50	Asym	1.000	0.043	0.005	1.000	0.045	0.004	1.000	0.040	0.007	1.000	0.044	0.010
	Sp Perm	1.000	0.052	0.005	1.000	0.050	0.004	1.000	0.046	0.002	1.000	0.051	0.001
	Perm	1.000	0.048	0.009	1.000	0.046	0.004	1.000	0.043	0.005	1.000	0.050	0.006
60×60	Asym	1.000	0.052	0.007	1.000	0.050	0.006	1.000	0.052	0.006	1.000	0.044	0.005
	Sp Perm	1.000	0.054	0.006	1.000	0.048	0.005	1.000	0.058	0.002	1.000	0.051	0.000
	Perm	1.000	0.053	0.010	1.000	0.045	0.008	1.000	0.051	0.005	1.000	0.048	0.004

block size is equal to $m = 10$.

Rejection rates based on a significance level of $\alpha = 0.05$ for all simulation settings are presented in Tables 1 and 2. Overall, all the test methods appear to maintain the expected rejection rates, which are 1.00 for H_{02}, 0.05 for H_{03}, and < 0.05 for H_{04}, based on $\alpha = 0.05$. Only for small samples sizes (30×30) did the asymptotic test show a rejection rate that is too small for kernel setting 2 and the skewed sample location pattern. Thus, for practical applications, smaller numbers of kernel functions might be preferable for the asymptotic test. For bootstrapping, the full spatial variants and those relying only on manipulating the hypothetical noise part perform equally well. Considering the computation time, the latter bootstrap variant might be preferable, as explored in Section 4.3.

4.2. Simulation study 2: Hypothesis testing for different signal and noise distributions

In these simulations, we compare the quality of the introduced tests for data distributions that are nonGaussian. To do so, we keep the same simulation outline and the same model settings as in the former section, but we consider a Gaussian and a nonGaussian distribution for the latent field. The latent field of the Gaussian setting (as in the former section) has a three-variate signal part and a two-variate standard normal noise part. The nonGaussian setting has a three-variate t-distributed signal part with degrees of freedom of 5, 6, and 7, and the two-variate noise part follows an exponential distribution (with zero mean and unit variance). The Gaussian and the nonGaussian settings have equal second-order spatial dependence, but the distributions are different; therefore, differences in the performance of the tests are the result of the different distributions. Moreover, we do not consider parametric bootstrap tests for these simulations,

Table 4. Rejection rates for model setting 1 with Gaussian and nonGaussian distributions and the skewed sample location pattern based on 2,000 simulation repetitions at a significance level of $\alpha = 0.05$.

		Kernel Setting 1						Kernel Setting 2					
		Gaussian			Non-Gaussian			Gaussian			Non-Gaussian		
Domain	Method	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}
30×30	Asym	1.000	0.047	0.004	1.000	0.035	0.007	1.000	0.040	0.004	1.000	0.051	0.005
	Sp Perm	1.000	0.060	0.004	1.000	0.046	0.007	1.000	0.066	0.001	1.000	0.074	0.001
	Perm	1.000	0.050	0.008	1.000	0.040	0.009	1.000	0.049	0.004	1.000	0.062	0.005
40×40	Asym	1.000	0.044	0.005	1.000	0.040	0.004	1.000	0.034	0.004	1.000	0.048	0.007
	Sp Perm	1.000	0.057	0.004	1.000	0.053	0.004	1.000	0.053	0.001	1.000	0.063	0.001
	Perm	1.000	0.051	0.007	1.000	0.048	0.007	1.000	0.040	0.004	1.000	0.053	0.006
50×50	Asym	1.000	0.043	0.004	1.000	0.043	0.002	1.000	0.044	0.006	1.000	0.040	0.008
	Sp Perm	1.000	0.054	0.006	1.000	0.054	0.002	1.000	0.058	0.002	1.000	0.060	0.001
	Perm	1.000	0.047	0.009	1.000	0.045	0.006	1.000	0.049	0.006	1.000	0.048	0.005
60×60	Asym	1.000	0.048	0.006	1.000	0.034	0.008	1.000	0.040	0.006	1.000	0.051	0.007
	Sp Perm	1.000	0.056	0.007	1.000	0.038	0.009	1.000	0.052	0.001	1.000	0.065	0.001
	Perm	1.000	0.048	0.011	1.000	0.039	0.011	1.000	0.046	0.004	1.000	0.057	0.004

Table 5. Rejection rates for model setting 2 with Gaussian and nonGaussian distributions and the uniform sample location pattern based on 2,000 simulation repetitions at a significance level of $\alpha = 0.05$.

		Kernel Setting 1						Kernel Setting 2					
		Gaussian			Non-Gaussian			Gaussian			Non-Gaussian		
Domain	Method	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}
30×30	Asym	1.000	0.040	0.003	1.000	0.044	0.007	1.000	0.046	0.005	1.000	0.050	0.005
	Sp Perm	1.000	0.044	0.003	1.000	0.049	0.006	1.000	0.054	0.001	1.000	0.062	0.001
	Perm	1.000	0.044	0.007	1.000	0.047	0.011	1.000	0.053	0.004	1.000	0.056	0.004
40×40	Asym	1.000	0.042	0.005	1.000	0.046	0.005	1.000	0.046	0.008	1.000	0.058	0.006
	Sp Perm	1.000	0.045	0.004	1.000	0.053	0.005	1.000	0.054	0.001	1.000	0.070	0.001
	Perm	1.000	0.044	0.008	1.000	0.048	0.006	1.000	0.051	0.006	1.000	0.060	0.004
50×50	Asym	1.000	0.050	0.004	1.000	0.046	0.006	1.000	0.053	0.007	1.000	0.048	0.005
	Sp Perm	1.000	0.050	0.004	1.000	0.050	0.004	1.000	0.061	0.002	1.000	0.053	0.000
	Perm	1.000	0.050	0.005	1.000	0.046	0.008	1.000	0.058	0.004	1.000	0.049	0.004
60×60	Asym	1.000	0.043	0.006	1.000	0.062	0.004	1.000	0.048	0.007	1.000	0.052	0.010
	Sp Perm	1.000	0.047	0.005	1.000	0.058	0.003	1.000	0.054	0.000	1.000	0.057	0.001
	Perm	1.000	0.046	0.008	1.000	0.057	0.007	1.000	0.051	0.004	1.000	0.052	0.004

because they are designed for Gaussian distributions.

Tables 3–6 summarize the rejection rates based on 2,000 simulation repetitions for a significance level of $\alpha = 0.05$ for model settings 1 and 2 with uniform and skewed coordinate patterns. These simulations again show the desired rejection rates of 1.00 for H_{02}, 0.05 for H_{03}, and < 0.05 for H_{04}, based on $\alpha = 0.05$. Most of the differences between the rejection rates for the Gaussian and nonGaussian cases are in the third decimal place. Thus the tests still performa well, even for heavy-tailed nonGaussian signal and noise distributions, and so we perform the subsequent simulations for the Gaussian case only.

Table 6. Rejection rates for model setting 2 with Gaussian and nonGaussian distributions and the skewed sample location pattern based on 2,000 simulation repetitions at a significance level of $\alpha = 0.05$.

		Kernel Setting 1						Kernel Setting 2					
		Gaussian			Non-Gaussian			Gaussian			Non-Gaussian		
Domain	Method	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}	H_{02}	H_{03}	H_{04}
30×30	Asym	1.000	0.042	0.007	1.000	0.040	0.004	1.000	0.038	0.005	1.000	0.043	0.004
	Sp Perm	1.000	0.050	0.007	1.000	0.048	0.002	1.000	0.051	0.002	1.000	0.066	0.001
	Perm	1.000	0.049	0.009	1.000	0.044	0.006	1.000	0.048	0.006	1.000	0.057	0.004
40×40	Asym	1.000	0.053	0.005	1.000	0.053	0.006	1.000	0.032	0.005	1.000	0.042	0.004
	Sp Perm	1.000	0.068	0.004	1.000	0.062	0.005	1.000	0.050	0.002	1.000	0.056	0.002
	Perm	1.000	0.058	0.010	1.000	0.052	0.006	1.000	0.040	0.005	1.000	0.050	0.005
50×50	Asym	1.000	0.040	0.004	1.000	0.051	0.004	1.000	0.040	0.007	1.000	0.048	0.004
	Sp Perm	1.000	0.051	0.004	1.000	0.055	0.002	1.000	0.054	0.002	1.000	0.064	0.001
	Perm	1.000	0.043	0.009	1.000	0.054	0.004	1.000	0.048	0.006	1.000	0.055	0.005
60×60	Asym	1.000	0.045	0.009	1.000	0.039	0.004	1.000	0.048	0.008	1.000	0.043	0.007
	Sp Perm	1.000	0.053	0.006	1.000	0.042	0.004	1.000	0.061	0.002	1.000	0.057	0.001
	Perm	1.000	0.048	0.010	1.000	0.040	0.006	1.000	0.053	0.006	1.000	0.044	0.004

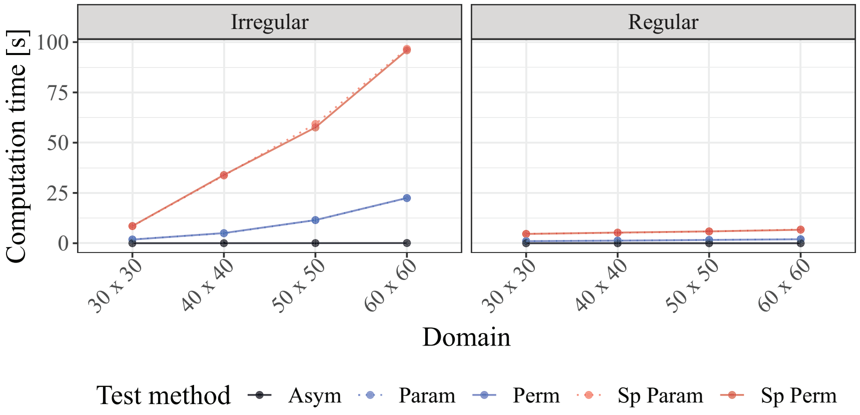

Figure 2. Median running times of the five test methods for different domain sizes with regular sampling sites based on five simulation repetitions. Computations are performed using code designed for regular and irregular sampling sites.

4.3. Simulation study 3: Computation time comparison

In this simulation, we investigate the computation times for the various test methods. As an illustrative example, we again consider a five-variate latent random field with model setting 1 and bivariate Gaussian noise components. In addition, we keep the same spatial domain sizes, although we change the sampling sites to be regular, defined as $[0, n] \times [0, n] \cap \mathbb{Z}^2$. H_{03} is tested using the same five test methods with the same number of bootstrap samples and block sizes. The key difference is that each test is performed using code designed for irregular sample locations, and code that considers simplifications made possible because the sample locations are regular (e.g., the simplified spatial bootstrap algorithm).

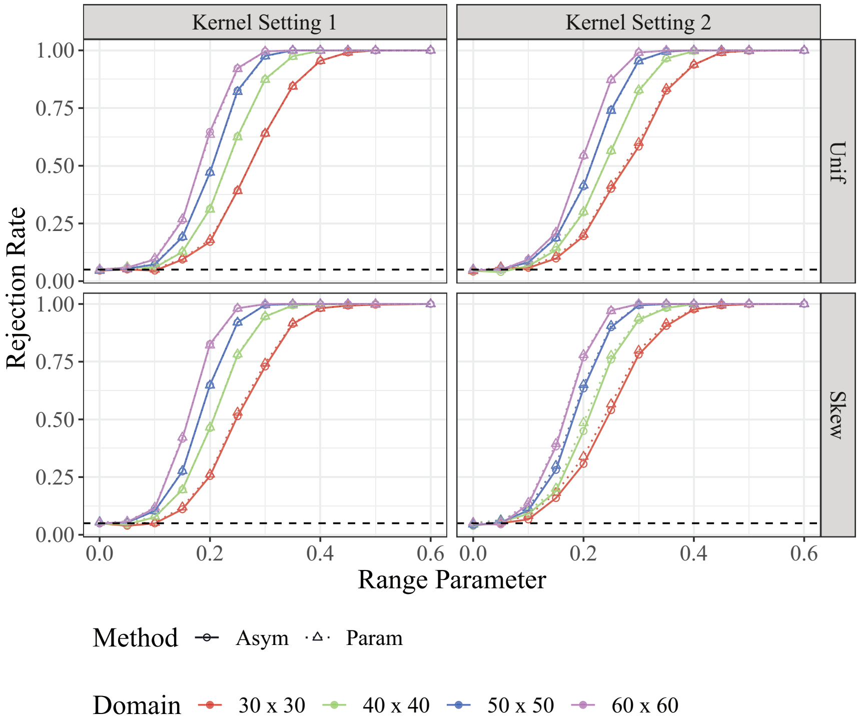

Figure 3. Rejection rates of the asymptotic and parametric bootstrap tests for H_{03} for different kernel settings as a function of the range parameter of the first entry of the signal part at a test significance level of 0.05 (indicated by the dashed line). The results are based on 2,000 repetitions.

We use two ring kernel functions with parameters $(r_1, r_2) \in \{(0,1),(1,\sqrt{2})\}$ for the irregular code, and kernels of the form $f(\mathbf{s}) = I(||\mathbf{s}|| = h)$ with $h \in \{1, \sqrt{2}\}$ for the regular code (one-way and two-way lag-1 local covariance matrices). This choice ensures that the same neighbors are selected for both versions of the code, and thus that the qualitative results of the tests are equal up to random effects of the bootstrap sampling procedures.

Figure 2 shows the median computation time based on five simulation repetitions carried out on a Windows machine with an Intel i5 CPU. The figure shows that the asymptotic tests are fastest, because the SBSS solution needs to be computed only once, whereas the bootstrap algorithms compute the SBSS solution B times.

Of greater interest is the overall difference in the computation time between regular and irregular code, possibly because the code for regular sampling sites does not rely on distances between sampling sites, as the irregular code does. Specifically, we can select the neighbors for the local covariance matrices by shifting the coordinate system appropriately for the regular code, whereas in the irregular code, this is based on looping over the distance matrix among all coordinates. This difference should also explain the different scaling of the computation time with increasing sample size, because looping through the distance matrix depends on the actual number of locations, whereas coordinate shifting does not.

Furthermore, there is a larger computation time difference between the full spatial bootstrap and the one that manipulates only the hypothetical noise for the irregular code compared with the regular one. This might be a result of the simplified spatial bootstrap variant for regular sampling sites. As explained above, for the irregular code, the distance matrix has to be computed for every new iteration, because the spatial bootstrap changes sampling sites for each iteration. In contrast, for the regular code, the sampling sites remain equal for each bootstrap iteration.

Overall, this simulation strongly indicates that regular sampling sites should be treated computationally as such. In addition, considering the overall similar performance of the tests in the former simulation, we can discard the spatial bootstrapping step for the irregular data, because it significantly increases the computation time.

4.4. Simulation study 4: Power of the test

In this part of the simulation, we investigate the power of the proposed tests. To do so, we keep the signal part of model setting 1 and the second entry of the noise (z_5) untouched, but replace the first entry of the noise part (z_4) with a signal following a Matérn correlation structure, with $\nu = 0.5$ and varying range parameter $\phi \in [0, 0.8]$. Note that the case $\phi = 0$ is technically forbidden in the Matérn covariance function; hence, we treat it simply as white noise. Expect for the case of $\phi = 0$, this setting has a true signal dimension $q = 4$; thus, we always test the wrong hypothesis H_{03}, and the test should be able to detect the true signal dimension more efficiently with an increasing range parameter. The hypothesis is tested using the asymptotic test and the parametric bootstrap test without full spatial bootstrapping because, although all tests in the hypothesis testing simulations performed similarly, these two test strategies showed a low computation time, which makes these simulations feasible.

Figure 3 depicts the test rejection rates at a significance level of $\alpha = 0.05$ as a function of the range parameter based on 2,000 simulation iterations for uniform and skewed sample location patterns. For lower sample sizes, all tests show a desired rejection rate of one at a range parameter of 0.5, which decreases to 0.3 when the sample size is highest. Interestingly, there are no differences between the skewed and uniform sample location patterns and the two kernel function settings considered. Furthermore, this simulation shows no significant difference between the asymptotic and the bootstrap tests, which again favors using the asymptotic test in practice.

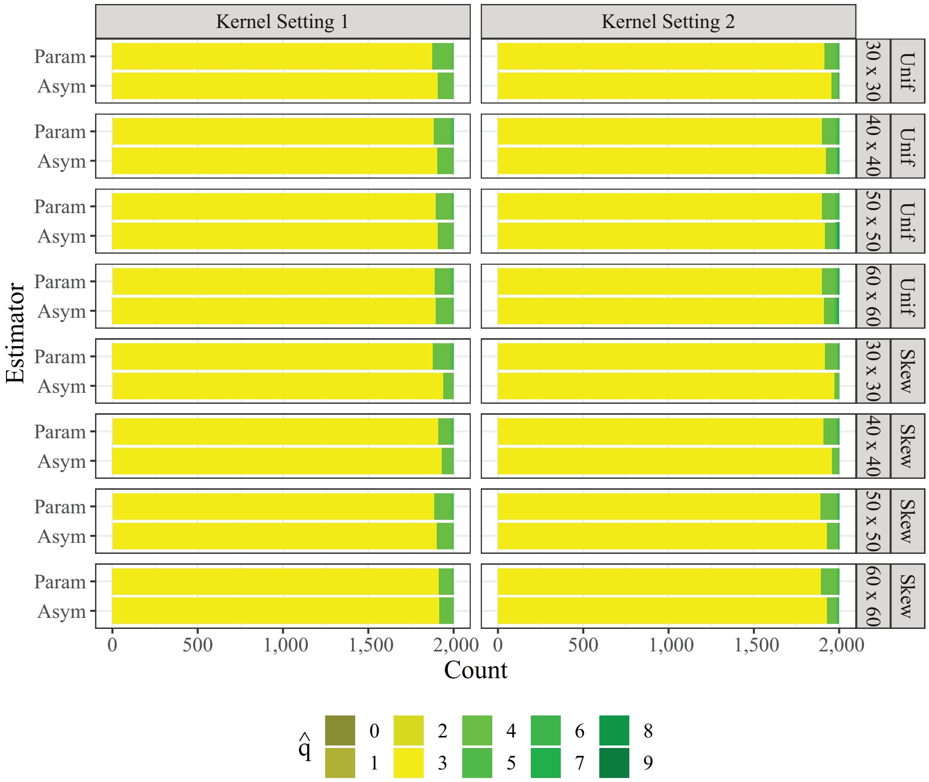

Figure 4. Frequencies of the estimated signal dimension for model setting 1.

4.5. Simulation study 5: Estimation of the signal dimension

The former simulations investigated only hypothesis tests for one specific value of the hypothetical signal dimension. In this section, we use hypothesis tests to estimate the signal dimension. We consider the same simulation settings as in Section 4.1, but increase the dimension of the noise part to seven, leading to a total latent random field dimension of $p = 10$, whereas the true signal dimension remains $q = 3$. We estimate the signal dimension using the divide-and-conquer strategy described above. As before, all hypothesis tests are performed using the asymptotic test method and the parametric bootstrap without full spatial bootstrapping. This choice is justified by the similar performance in terms of signal dimension testing of all bootstrap test variants, and the fact that the full spatial bootstrap is computationally unfeasible for such a large simulation.

Figures 4 and 5 depict the estimated dimensions for 2,000 simulation repetitions for a significance level of $\alpha = 0.05$. Overall, the estimation is highly accurate, with the estimated dimension being equal to the true dimension in approximately 95% of the cases. Interestingly, the signal dimension is never underestimated, but is overestimated in approximately 100 of the simulation

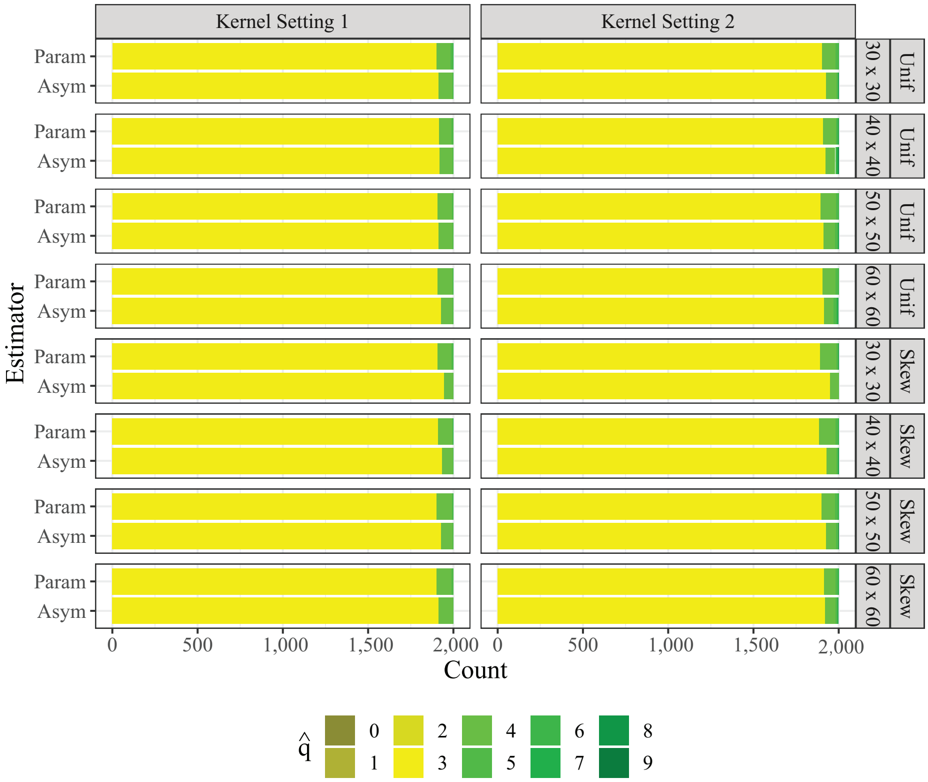

Figure 5. Frequencies of the estimated signal dimension for model setting 2.

iterations, reflecting the significance level $\alpha = 0.05$. For all settings, the asymptotic test outperformed the bootstrap test, particularly for low sample sizes, which is a counter-intuitive result. This may be because, as the former simulations show, for low sample sizes, the asymptotic test never met the theoretical rejection rate, which is simply the significance level when the null is actually true for small sample sizes (Tables 1 and 2). Therefore, the true null is accepted more often, leading to better performance when estimating the signal dimension.

5. Conclusion

In this paper, we have proposed testing and estimation methods for the number of latent signal components in the SBSS model. The asymptotic null distributions of the test statistic are given under various conditions, without assuming the domain is necessarily regular. A consistent estimator of the dimension based on the sequential tests is also introduced. For small sample cases, different bootstrap strategies are suggested. In addition to the theoretical results, the five simulation studies presented in Section 4 demonstrate that our asymptotic tests are comparable with bootstrap tests in terms of hypothesis testing and estimation. In terms of computation time, our asymptotic method is much faster than the bootstrap tests. When a regular domain structure is used, the computation time can be decreased significantly.

Our proposed dimension tests in the SBSS context might be useful for further analysis of the latent fields, including for various forms of spatial prediction. Indeed, the components of the latent field are uncorrelated, and thus predictions can be carried out on each latent field independently. This leads to a reduction as a result of building several models, rather than a single multivariate model. This procedure has been investigated and found to be useful by Muehlmann, Nordhausen and Yi (2021). As an additional step, one of our proposed dimension tests can be perfomed before the spatial prediction, leading to a reduction of the latent field dimension, resulting in even fewer univariate models needing to be built.

However, it is not clear how to obtain the sequence mentioned in Proposition 4 in a data-driven way, leading to a consistent estimate. In future research, we plan to develop a ladle estimator (Luo and Li (2016, 2021)) for this setting that will be based either on bootstrapping or on data augmentation. Other ideas for future research are to develop similar approaches for spatiotemporal data and to study the fixed-domain asymptotic properties (Stein (1995); Cressie (1993, Sec. 5.8)) of SBSS. It may be interesting to study the SBSS model in a high-dimension framework, where we could transfer SBSS to a spiked model when both n and p go to infinity. The problem of identifying the number of spikes is studied in, for example, Passemier and Yao (2014). When $p < n$ is diverging, Zhang, Hao and Yao (2022) propose a new way of estimating the mixing matrix for an SBSS model. Thus, combining the two methods might provide insight into selecting the high-dimensional signal. However, we suspect that it will be difficult to investigate the limiting behavior of eigenvalues in a spatial setting.

Supplementary Material

The Supplementary Material contains all technical proofs, as well as an environmental data example.

Acknowledgments

The authors would like to thank the editor, associate editor, and two anonymous reviewers for their constructive comments and suggestions in the review process. The work of Christoph Muehlmann, Klaus Nordhausen, and Mengxi Yi was supported by the Austrian Science Fund (No. P31881-N32). The work of Mengxi Yi was also supported by the National Natural Science Foundation of China (No. 12101119).

References

Bachoc, F., Betancourt, J., Furrer, R. and Klein, T. (2020a). Asymptotic properties of the maximum likelihood and cross validation estimators for transformed Gaussian processes. *Electronic Journal of Statistics* **14**, 1962–2008.

Bachoc, F., Genton, M. G., Nordhausen, K., Ruiz-Gazen, A. and Virta, J. (2020b). Spatial blind source separation. *Biometrika* **107**, 627–646.

Bevilacqua, M., Gaetan, C., Mateu, J. and Porcu, E. (2012). Estimating space and space-time covariance functions for large data sets: A weighted composite likelihood approach. *Journal of the American Statistical Association* **107**, 268–280.

Bivand, R. S., Pebesma, E. and Gomez-Rubio, V. (2013). *Applied Spatial Data Analysis with R.* 2nd Edition. Springer, New York.

Bodenham, D. A. and Adams, N. M. (2016). A comparison of efficient approximations for a weighted sum of Chi-squared random variables. *Statistics and Computing* **26**, 917–928.

Bura, E. and Cook, R. D. (2001). Extending sliced inverse regression. *Journal of the American Statistical Association* **96**, 996–1003.

Chernick, M. R., González-Manteiga, W., Crujeiras, R. M. and Barrios, E. B. (2011). *Bootstrap Methods*, 169–174. Springer, Berlin, Heidelberg.

Comon, P. and Jutten, C. (2010). *Handbook of Blind Source Separation: Independent Component Analysis and Applications.* Academic Press, Amsterdam.

Cressie, N. (1993). *Statistics for Spatial Data.* John Wiley & Sons.

De Iaco, S., Myers, D., Palma, M. and Posa, D. (2013). Using simultaneous diagonalization to identify a space–time linear coregionalization model. *Mathematical Geosciences* **45**, 69–86.

Dudley, R. M. (2018). *Real Analysis and Probability.* CRC Press.

Emery, X. (2010). Iterative algorithms for fitting a linear model of coregionalization. *Computational Geosciences* **36**, 1150–1160.

Genton, M. G. and Kleiber, W. (2015). Cross-covariance functions for multivariate geostatistics. *Statistical Science* **30**, 147–163.

Gneiting, T., Kleiber, W. and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association* **105**, 1167–1177.

Goulard, M. and Voltz, M. (1992). Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix. *Mathematical Geology* **24**, 269–.

Gräler, B., Pebesma, E. and Heuvelink, G. (2016). Spatio-Temporal Interpolation using gstat. *The R Journal* **8**, 204–218.

Hijmans, R. J. (2020). *raster: Geographic Data Analysis and Modeling.* R package version 3.1-5.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data.* Springer, New York.

Legendre, P. and Legendre, L. F. (2012). *Numerical Ecology.* Elsevier.

Luo, W. and Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika* **103**, 875–887.

Luo, W. and Li, B. (2021). On order determination by predictor augmentation. *Biometrika* **108**, 557–574.

Matilainen, M., Nordhausen, K. and Virta, J. (2018). On the number of signals in multivariate time series. In *Latent Variable Analysis and Signal Separation*, 248–258. Springer International Publishing, Cham.

Miettinen, J., Nordhausen, K. and Taskinen, S. (2017). Blind source separation based on joint diagonalization in R: The packages JADE and BSSasymp. *Journal of Statistical Software* **76**, 1–31.

Muehlmann, C., Bachoc, F. and Nordhausen, K. (2022). Blind source separation for non-stationary random fields. *Spatial Statistics* **47**, 100574.

Muehlmann, C., Nordhausen, K. and Virta, J. (2020). *SpatialBSS: Blind Source Separation for Multivariate Spatial Data*. R package version 0.8.

Muehlmann, C., Nordhausen, K. and Yi, M. (2021). On cokriging, neural networks, and spatial blind source separation for multivariate spatial prediction. *IEEE Geoscience and Remote Sensing Letters* **18**, 1931–1935.

Nordhausen, K. and Oja, H. (2018). Independent component analysis: A statistical perspective. *WIREs: Computational Statistics* **10**, e1440.

Nordhausen, K., Oja, H., Filzmoser, P. and Reimann, C. (2015). Blind source separation for spatial compositional data. *Mathematical Geosciences* **47**, 753–770.

Nordhausen, K., Oja, H. and Tyler, D. E. (2022). Asymptotic and bootstrap tests for subspace dimension. *Journal of Multivariate Analysis* **188**, 104830.

Nordhausen, K., Oja, H., Tyler, D. E. and Virta, J. (2017). Asymptotic and bootstrap tests for the dimension of the non-gaussian subspace. *IEEE Signal Processing Letters* **24**, 887–891.

Nordhausen, K. and Ruiz-Gazen, A. (2022). On the usage of joint diagonalization in multivariate statistics. *Journal of Multivariate Analysis* **188**, 104844.

Nordhausen, K. and Virta, J. (2018). Ladle estimator for time series signal dimension. In *2018 IEEE Statistical Signal Processing Workshop (SSP)*, 428–432.

Nordman, D. J., Lahiri, S. N. and Fridley, B. L. (2007). Optimal block size for variance estimation by a spatial block bootstrap method. *Sankhyā* **69**, 468–493.

Passemier, D. and Yao, J. (2014). Estimation of the number of spikes, possibly equal, in the high-dimensional case. *Journal of Multivariate Analysis* **127**, 173–183.

R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Schlather, M., Malinowski, A., Menck, P. J., Oesting, M. and Strokorb, K. (2015). Analysis, simulation and prediction of multivariate random fields with package RandomFields. *Journal of Statistical Software* **63**, 1–25.

Schmidt, A. M. and Gelfand, A. E. (2003). A Bayesian coregionalization approach for multivariate pollutant data. *Journal of Geophysical Research: Atmospheres* **108**.

Shaby, B. and Ruppert, D. (2012). Tapered covariance: Bayesian estimation and asymptotics. *Journal of Computational and Graphical Statistics* **21**, 433–452.

Stein, M. L. (1995). Fixed-domain asymptotics for spatial periodograms. *Journal of the American Statistical Association* **90**, 1277–1288.

Virta, J. and Nordhausen, K. (2021). Determining the signal dimension in second order source separation. *Statistica Sinica* **31**, 135–156.

von Storch, H. and Zwiers, F. W. (2001). *Statistical Analysis in Climate Research*. Cambridge University Press.

Wackernagel, H. (1994). Cokriging versus kriging in regionalized multivariate data analysis. *Geoderma* **62**, 83–92.

Wackernagel, H. (2003). *Multivariate Geostatistics*. Springer.

Zhang, B., Hao, S. and Yao, Q. (2022). Blind source separation over space: An eigenanalysis approach. LSE Research Online Documents on Economics 121093. London School of Economics and Political Science, London.

Christoph Muehlmann

Computational Statistics, Vienna University of Technology, 1040 Vienna, Austria.

E-mail: christoph.muehlmann@tuwien.ac.at

François Bachoc

Institut de Mathématiques de Toulouse, Université Paul Sabatier, 31062 Toulouse, France.

E-mail: francois.bachoc@math.univ-toulouse.fr

Klaus Nordhausen

Computational Statistics, Vienna University of Technology, 1040 Vienna, Austria.
Department of Mathematics and Statistics, University of Jyväskylä, 40014 Jyväskylä, Finland.

E-mail: klaus.k.nordhausen@jyu.fi

Mengxi Yi

School of Statistics, Beijing Normal University, 100875 Beijing, China.
Computational Statistics, Vienna University of Technology, 1040 Vienna, Austria.

E-mail: mxyi@bnu.edu.cn

(Received September 2021; accepted September 2022)