Statistica Sinica Preprint No: SS-2021-0326	
Title	Test of the Latent Dimension of a Spatial Blind Source Separation Model
Manuscript ID	SS-2021-0326
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0326
Complete List of Authors	Christoph Muehlmann, Francois Bachoc, Klaus Nordhausen and Mengxi Yi
Corresponding Authors	Mengxi Yi
E-mails	mengxi.yi@outlook.com
Notice: Accepted version subject to English editing.	

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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 is 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 in the Supplemental Material demonstrate that our test is at least comparable to and often outperforms bootstrap-based techniques, which are also introduced in this paper.

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

1. Introduction

With the advance of technology, massive amounts of multivariate spatial data can be collected. As one example, researchers may use these datasets to investigate various issues in geographical, ecological (Legendre and Legendre, 2012) or atmospheric (Von Storch and Zwiers, 2001) sciences. From a domain experts perspective proper analysis of such multivariate spatial data is carried out by at least investigating and interpreting p maps (for the p measured variables) which might be con-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

taminated by various sources of noise, such as measurement inconsistencies or errors. Moreover, the interpretation of the raw data might be complicated as the original variable reflect a mixture of physical processes which are actually of interest. As such also dependencies between measurements need to be investigated. From a statisticians perspective these spatially correlated datasets contain dependencies both within and among the individual data processes, which makes statistical modeling of the multivariate spatial data a challenge. The difficulty of modeling is further intensified when the dimensionality p is large. With a dataset of size n, it takes a total of $cp(p+1)/2$ 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. Further, 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, Section 3) and 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. (2020). 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 $\mathbf{\Gamma}$, i.e. $\mathbf{\Omega}^{-1}$, Nordhausen et al. (2015) have proposed an estimator based on the simultaneous diagonalization of two scatter matrices, and Bachoc et al. (2020) have extended this method to jointly diagonalize more than two scatter matrices for multivariate spatial data. Pre-processing the data with such a SBSS method is appealing from a

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

practitioners perspective as the latent components more likely reflect the physical nature of the processes that generated the data. For example Nordhausen et al. (2015) found six physical meaningful latent components in a geostatistical dataset which were not easily detectable in the original data. Moreover, it suffices to investigate only p maps as the resulting latent components are spatially uncorrelated. From a statisticians perspective common tasks such as modeling of the spatial covariance or predictions of the original data are again modified as the statistical analysis can be carried out with univariate tools on the latent components. The analysis results of the latent components can be simply transferred back to the original data by using the fact that the transformation is linear in its nature. Muehlmann et al. (2021) investigate this procedure in the context of geostatistical prediction. Avoiding the task of building one multivariate model in favor of p univariate ones simplifies the given tasks significantly. Nevertheless, if the dimension p is still high a further reduction is desirable. This reduction can be obtained by the fact that not all of the p components might be of interest.

The SBSS model of Nordhausen et al. (2015) gives no preference to any of the latent components, with all p of them being basically 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 to be regarded as the signal, while the remaining components are discarded as noise. This can be translated 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 et al. (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 less components which need to be investigated by practitioners and modeled by statisticians which in turn simplifies the desired analysis.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

In this paper, we consider SBSS in which signals are characterized as components having 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 to extend the framework of Bachoc et al. (2020), to the case where the signal and noise components are not all asymptotically identifiable, as well as where their distributions are not necessarily Gaussian. We develop new proof techniques to obtain these two extensions. In particular, the first extension is based on generalizing arguments made by Virta and Nordhausen (2021) to a spatial setting. The second one is based on extending arguments in Bachoc et al. (2020) beyond the case of transformed Gaussian processes.

In addition, we demonstrate that introducing new scatter matrices compared to the one used by Bachoc et al. (2020) enables the obtainment of 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 drastic computational cost reduction for subsequent multivariate spatial modeling where then only the signal components are used.

We put forward several bootstrap versions of the test. For both the asymptotic and bootstrap tests, we demonstrate that computational gains are obtained when the observation locations are gridded. In an extensive simulation study, we then show that the various tests already have levels close to the nominal one, for small to moderate sample sizes. We also observe an accurate estimation of the signal dimension. We conclude that the asymptotic test is comparable to and often outperforms the bootstrap ones while being less computationally demanding. Employing an environmental application, we then show that our methods enable the reduction of the dimen-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

sion 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 then described in Section 3 and the simulation results are reported in Section 4. We finally discuss some concluding remarks in Section 5. The proofs of the theoretical results and the environmental application are presented in the Supplemental Material.

2. Setup and Model

Suppose our data consists of a p-dimensional multivariate random field $\mathbf{x}(\mathbf{s}) = \{x_1(\mathbf{s}), \cdots, x_p(\mathbf{s})\}^T$, $\mathbf{s} \in \mathcal{S}$, 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. We can refer, for instance, to De Iaco et al. (2013), Genton and Kleiber (2015) and Gneiting et al. (2010) for an introduction and various approaches to modeling these.

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

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

where $\mathbf{\Omega}$ is a $p \times p$ unknown invertible matrix, and $\mathbf{z}(\mathbf{s}) = \{z_1(\mathbf{s}), \cdots, z_p(\mathbf{s})\}^T$ is the latent field having independent components with $\mathrm{Cov}(\mathbf{z}(\mathbf{s})) = \mathbf{I}_p$ for all $\mathbf{s} \in \mathcal{S}$. It is interesting to see the connection of the SBSS model to one very popular multivariate covariance model, namely the linear model of coregionalization (LMC) which writes as

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

Here, $\mathbf{T}_m$ are non-negative definite $p \times p$ coregionalization matrices and $\rho_m(h)$ are univariate stationary correlation functions. Details on the LMC can be found in for example Goulard and Voltz

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

(1992); Schmidt and Gelfand (2003); Emery (2010). Dimension reduction in the LMC literature is carried out by firstly 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 an eigendecomposition of the coregionalization matrices. If the system of eigenvectors is equal across a few summands this hints that these matrices are proportional. This is referred to as intrinsic correlation, details are provided by Wackernagel (1994). The latter is addressed by Goulard and Voltz (1992) which describe that the coregionalization matrices arise from a scalar product matrix of latent variables. Variants of a principal component analysis (PCA) of the coregionalization matrices lead to a lower dimensional representation of these latent variables. This is denoted as regionalized PCA, see also (Wackernagel, 2003, Chapter 27) for details.

As pointed out by Bachoc et al. (2020) 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 m-th column of the mixing matrix $\boldsymbol{\Omega}$) leading to rank-one coregionalization matrices and the $\rho_m(h)$ correspond to 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) is done without estimating or specifying a model for the covariances of the latent field components. Moreover, our approach in dimension reduction is different in the sense that we test if some latent components are white noise. This leads to a reduction of r.

Next we present how to estimate the unmixing matrix $\boldsymbol{\Gamma}$, i.e. $\boldsymbol{\Omega}^{-1}$, 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 $f_0, f_1, \cdots, f_k$, with $f_\ell : \mathbb{R}^d \to \mathbb{R}$ for $\ell = 0, \cdots, k$, and with $f_0(\mathbf{s}) = I(\mathbf{s} = \mathbf{0})$. Note that we call $f_0, f_1, \cdots, f_k$ kernels, similarly as the past references Bachoc et al. (2020); Muehlmann et al. (2022), and for instance analogously to kernel smoothing,

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

but $f_0, f_1, \cdots, f_k$ should not be confused with the covariance functions of the components of $\mathbf{x}$ or $\mathbf{z}$. For $f \in \{f_0, f_1, \cdots, f_k\}$, 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. Notice that $F_{n,f_0} = 1$. Let $f \in \{f_1, \cdots, f_k\}$. 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. (2020), where simply n is used. Here, the introduction of $F_{n,f}^{1/2}$ 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 $\widehat{\mathbf{M}}(f_0), \widehat{\mathbf{M}}(f_1), \cdots, \widehat{\mathbf{M}}(f_k)$ are used to estimate the unmixing matrix $\mathbf{\Gamma}$ as

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

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

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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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, $\sum_{\ell=1}^{k} \widehat{\mathbf{D}}^2_{\ell,1,1} \geqslant \cdots \geqslant \sum_{\ell=1}^{k} \widehat{\mathbf{D}}^2_{\ell,p,p}$. We are now interested in the case in which there are q "real" continuous random fields in $\mathbf{z}$, while the remaining $p - q$ components are white noise.

For $q \in \{0, \cdots, p-1\}$, 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 $a = 1, \cdots, p-q$, 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 $\ell = 1, \cdots, k$, f_ℓ is symmetric and satisfies $f_\ell(\mathbf{0}) = 0$. For $a = 1, \cdots, q$, 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 come first and are followed by the $p - q$ noise components. As 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 $z_1, \cdots, z_q$ is sufficient for these signals to be asymptotically separated from the noise fields $z_{q+1}, \cdots, z_p$. It should also be noted that we do not need to consider the stronger assumption that the q vectors of the a-th diagonal elements in $\mathbf{\Omega}^{-1}\mathbf{M}(f_1)\mathbf{\Omega}^{-T}, \cdots, \mathbf{\Omega}^{-1}\mathbf{M}(f_k)\mathbf{\Omega}^{-T}$, for $a = 1, \cdots, q$, for the signal random fields, are asymptotically distinct (see Assumption 9 in Bachoc et al. (2020)) and non-zero.

We remark that when Condition 2 is satisfied by $f_1, \ldots, f_k$, it is likely to be also satisfied with the single kernel $f_1 + \cdots + f_k$. This means that using a single kernel can be sufficient to

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

obtain the various asymptotic results of Section 3 on the test statistic 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 individually some of the first (most important) signal components. As shown in Bachoc et al. (2020), this usually requires multiple kernels, both for theoretical guarantees and practical efficiency. Using the same set of kernels for these two studies (signal dimension and components) can be desirable for the user, for instance for interpretability reasons. For more details on joint diagonalization in multivariate methods see also Nordhausen and Ruiz-Gazen (2022).

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

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

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 then 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 the dataset provides indications that more than q signal components are present. For a nominal level $\alpha \in (0,1)$, the threshold will be set to the quantile $1-\alpha$ of the asymptotic distribution of Proposition 1 or 2 or Corollary 1, depending on the context.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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 $z_1, \cdots, z_p$ are independent, centered and stationary.*

The independence assumption makes the study of 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 computations of the proof of Proposition 1. The stationarity assumption is standard in spatial statistics, see for instance Shaby and Ruppert (2012); Bachoc et al. (2020). The zero mean assumption is replaced by a constant unknown mean assumption in Section 3.4. For $a = 1, \cdots, p$, we let z_a have 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, for all $i \neq j, i, j = 1, \cdots, n, ||\mathbf{s}_i - \mathbf{s}_j|| \geqslant \delta$.*

Condition 4 implies that we are dealing with the increasing-domain asymptotic framework. For examples, see Cressie (1993, Section 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 $a = 1, \cdots, q$, for $u, v \in \mathbb{N}$, $u \geqslant 1$, $v \geqslant 1$, $u + v \leqslant 4$, for $\mathbf{y}_1, \ldots, \mathbf{y}_u \in \mathbb{R}^d$, for $\mathbf{w}_1, \ldots, \mathbf{w}_v \in \mathbb{R}^d$,*

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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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

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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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*

$$\max_{a=1,\ldots,p-q} \sup_{\mathbf{y}\in\mathbb{R}^d} \mathbb{E}\left(|z_{q+a}(\mathbf{y})|^{4+\alpha}\right) < \infty. \tag{3.6}$$

Condition 6 requires the noise values to 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 the use of Lemma 6 in the Supplemental Material) and to obtain a central limit theorem (see Lemma 4 in the Supplemental Material). The condition in (3.6), 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 $\ell = 1, \cdots, k$*

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

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

$$R(r_1, r_2)(\mathbf{s}) = I(r_1 < ||\mathbf{s}|| \leqslant r_2), \tag{3.7}$$

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

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

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

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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

compared with n.

Condition 9. *For all* $\ell, \ell' = 1, \cdots, k, \ell \neq \ell', f_\ell(\mathbf{y}) f_{\ell'}(\mathbf{y}) = 0$ *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 where 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. (2020) for the local covariance matrices, and when 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*

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$\widehat{\mathbf{M}}(f)$ *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 $f \in \{f_1, \cdots, f_k\}$. *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* $(\mathbf{V}_{\ell,a,b})_{\ell=1,\cdots,k,a,b=1,\cdots,p-q}$ *is a Gaussian vector with mean vector* $\mathbf{0}$ *and with 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 $\ell, \ell' = 1, \cdots, k$ *and* $a, b, a', b' = 1, \cdots, p-q$. *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 e.g., Bodenham and Adams (2016) for a presentation of 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 $\mathcal{X}_1^2, \cdots, \mathcal{X}_k^2$ *are independent and are chi-squared distributed 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, i.e., $\mathcal{S} \subseteq \mathbb{Z}^d$, the kernel functions can be based on the natural notion of a spatial neighborhood on the grid, which simplifies our technique.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

A location $\mathbf{s}_0 = (s_1, \cdots, s_d) \in \mathbb{Z}^d$ has $2d$ one-way lag-h neighbors, $(s_1 \pm h, \cdots, s_d), (s_1, s_2 \pm h, \cdots, s_d), \cdots, (s_1, \cdots, s_{d-1}, s_d \pm h)$. For example, if $d = 2$ and $h = 1$, the 4 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 could define the one-way lag-1 population and sample local covariance matrices as

$$\mathbf{M} = \frac{1}{\sqrt{n \sum_{i=1}^n |\mathcal{N}_{\mathbf{s}_i}|}} \sum_{i=1}^n \sum_{\mathbf{s}_j \in \mathcal{N}_{\mathbf{s}_i}} \mathbb{E}\left(\mathbf{x}(\mathbf{s}_i)\mathbf{x}(\mathbf{s}_j)^T\right)$$
$$\text{and } \widehat{\mathbf{M}} = \frac{1}{\sqrt{n \sum_{i=1}^n |\mathcal{N}_{\mathbf{s}_i}|}} \sum_{i=1}^n \sum_{\mathbf{s}_j \in \mathcal{N}_{\mathbf{s}_i}} \mathbf{x}(\mathbf{s}_i)\mathbf{x}(\mathbf{s}_j)^T \tag{3.8}$$

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 4 two-way lag-1 neighbors that are of the form $(s_1 \pm 1, s_2 \pm 1)$. In general, for $m, h \in \mathbb{N}$, $1 \leqslant m \leqslant d$, the m-way lag-h population and sample local covariance matrices can be defined as:

$$\mathbf{M} = \frac{1}{\sqrt{n \sum_{i=1}^n |\mathcal{N}^m_{h,\mathbf{s}_i}|}} \sum_{i=1}^n \sum_{\mathbf{s}_j \in \mathcal{N}^m_{h,\mathbf{s}_i}} \mathbb{E}\left(\mathbf{x}(\mathbf{s}_i)\mathbf{x}(\mathbf{s}_j)^T\right)$$
$$\text{and} \quad \widehat{\mathbf{M}} = \frac{1}{\sqrt{n \sum_{i=1}^n |\mathcal{N}^m_{h,\mathbf{s}_i}|}} \sum_{i=1}^n \sum_{\mathbf{s}_j \in \mathcal{N}^m_{h,\mathbf{s}_i}} \mathbf{x}(\mathbf{s}_i)\mathbf{x}(\mathbf{s}_j)^T, \tag{3.9}$$

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

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

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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

In general, in Equations (2.2) and (2.3) the m-way lag-h population and sample local covariance matrices $\mathbf{M}$ and $\widehat{\mathbf{M}}$ can also be written in the form $\mathbf{M}(f)$ and $\widehat{\mathbf{M}}(f)$, 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, the same limiting conclusions can be derived. 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. Now we wish to test the null hypothesis, for $r \in \{0, \cdots, p-1\}$,

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

This hypothesis states that the signal dimension is r. Similar 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 $t_r, r = 0, 1, \cdots, p-1$ 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,*

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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

A consistent estimate $\hat{q}$ of the unknown signal dimension $q \leqslant p - 1$ 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*

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

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

Specifying the sequence c_n is however not obvious in practice and still an open problem in order determination using hypothesis tests based on eigenvalues as can be done similarly in principal component analysis (PCA), sliced inverse regression or independent components analysis, see for example Bura and Cook (2001); Nordhausen et al. (2017, 2022) and 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 will 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 of Corollary 1 are unchanged when

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$\mathbf{z}$ has a non-zero unknown constant mean function and when the observations are empirically centered for the computation of the local covariance matrices.

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

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

$$\textit{and} \quad \mathbf{M}(f_0) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}(\mathbf{s}_i) - \bar{\mathbf{x}})(\mathbf{x}(\mathbf{s}_i) - \bar{\mathbf{x}})^T, \tag{3.10}$$

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 as well as that of Corollary 1 still hold under the same assumptions, except that $\widehat{\mathbf{M}}(f)$ *is everywhere replaced by* $\overline{\mathbf{M}}(f)$.

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 a large sample size may be needed for the finite sample level to match the asymptotic one. As an alternative for smaller sample sizes, we can formulate noise dimension tests based on the bootstrap.

In its original form, the bootstrap is a non-parametric tool for estimating the distribution of an estimator or test statistic by re-sampling from the empirical cumulative distribution function (ECDF) of the sample at hand. It has had good performance in many statistical problems by theoretical analysis as well as simulation studies and applications to real data. See Chernick et al. (2011) or Lahiri (2003) for a more detailed discussion.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Again, we assume that the observed random field is following 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 seen in Equation (2.5). In the following, we formulate a method for re-sampling from the distribution of Model (2.1) by respecting the null hypothesis H_{0r}. In line with the ideas presented by Matilainen et al. (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 manipulating only the hypothetical noise parts $(\hat{\mathbf{z}}(\mathbf{s}))_i$ for $i = r+1, \ldots, p$ and all $\mathbf{s} \in \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ in one of the following ways.

Parametric: Here, it is assumed that each noise part is independent and identically distributed (iid) 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 iid 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 $\hat{\mathbf{z}}$.

After replacing the hypothetical noise part by a bootstrap sample in one of the former ways, the goal of sampling from Model (2.1) under H_{0r} is achieved. However, so far the uncertainty of estimating the signal has not been considered in the bootstrap test. Therefore, an optional second step in the whole re-sampling procedure is devoted to drawing a spatial bootstrap sample from the already manipulated sample as follows. We suggest the application of spatial bootstrapping as discussed in Lahiri (2003), and in the following we summarize the main ideas. Let us 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 non-overlapping blocks

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

of size m^d that lie partially in $\mathcal{S}$, formally $\mathcal{B} = \{b_{\mathbf{i}} = (\mathbf{i} + (0,1]^d)m \cap \mathcal{S} : (\mathbf{i} + (0,1]^d)m \cap \mathcal{S} \neq \varnothing, \mathbf{i} \in \mathbb{Z}^d\}$, 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 $b_{\mathbf{j}}$, while the shape changes to that of $b_{\mathbf{i}}$, taking care of the boundary blocks that do not fully lie within $\mathcal{S}$. Finally, the bootstrapped version of the random field writes 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, as it relies on the distances between all sampling sites, which need to be newly computed in each iteration. For regular data, this can be avoided by using a slightly different bootstrap regime as follows.

Nordman et al. (2007) have suggested a slightly different approach for sampling sites located on a regular grid, meaning that the sampling sites satisfy $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subseteq \mathcal{S} \cap \mathbb{Z}^d$. Again, the domain $\mathcal{S}$ is divided into blocks of size m^d that are either non-overlapping 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 by $\{\mathbf{z}(\mathbf{s}) : \mathbf{s} \in b_{\mathbf{j}} \cap \mathbb{Z}^d\}$ 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 local covariance matrices, as only the random field values change. We will compare the computation times of the former two approaches in the simulation

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

study presented in Section 4.3.

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 test statistic $t = t_r(\mathbf{X})$;

for $k \in \{1, \ldots, B\}$ **do**

  Replace the last $p - r$ columns of $\hat{\mathbf{Z}}$ by 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: $[\#(t^k \geqslant t) + 1]/(B + 1)$;

Algorithm 2 Divide and Conquer

Set $lower$, $upper$ and α;

$middle = \lfloor (upper - lower)/2 \rfloor$;

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 $\hat{q} = middle + 1$;

Algorithm 1 summarizes the formerly discussed bootstrap strategy to test for one specific value of signal dimension r. To estimate the signal dimension, a sequence of tests for different signal dimensions r at a given significance level α are carried out. A number of different test sequences are possible, but we rely on a divide-and-conquer strategy outlined in Algorithm 2. Here, the $test_function$ could be either one of the bootstrap test variants seen in Algorithm 1 or the asymp-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

totic test outlined above.

4. Simulation

To validate the performance of the methods introduced above, 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 et al. (2020), `JADE` from Miettinen et al. (2017), `sp` from Bivand et al. (2013), `raster` from Hijmans (2020), `gstat` from Gräler et al. (2016) and `RandomFields` from Schlather et al. (2015).

4.1 Simulation Study 1: Hypothesis Testing

In this part of the simulation, we explored the performance of hypothesis testing. For all the following simulations, we considered the SBSS model, as shown in Equation (2.1), where without loss of generality we set $\mu_a = 0$ for $a = 1, \dots, p$ and assume the mean to be unknown. For the latent signal part we used two different three-variate random field model settings. Therefore, the true dimension is always $q = 3$. All the random fields were Gaussian and followed a Matérn correlation structure, and the a-th random field z_a thus had 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 second kind with shape parameter ν, Γ is the gamma function. The parameters used were $(\nu, \phi) \in \{(3, 2), (2, 1.5), (1, 1)\}$ and $\{(3, 2), (2, 1.5), (0.6, 0.6)\}$ for model setting 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 iid samples drawn from $N_2(\mathbf{0}, \mathbf{I}_2)$,

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 1: Rejection rates for model setting 1 based on 2000 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}
	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
30×30	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
	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
40×40	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
	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
50×50	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
	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
60×60	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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 2: Rejection rates for model setting 2 based on 2000 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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

leading to a total latent field dimension of $p = 5$ for both model settings. As SBSS is affine equivariant (for details see Bachoc et al. (2020) and the Supplement) we chose the mixing matrix to be the identity matrix, i.e., $\mathbf{\Omega} = \mathbf{I_5}$, without loss of generality.

We focused 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 considered two different sample location patterns: uniform and skewed. For the uniform pattern, n^2 pairs of (x, y)-coordinates were randomly drawn from a uniform distribution $U(0, 1)$ and then multiplied by n, leading to a constant sampling location density over the entire domain. We followed the same approach for the skewed pattern, with the only difference being that the x coordinate values were drawn from a beta distribution $\beta(2, 5)$, resulting in a denser arrangement of samples in the left half of the domain.

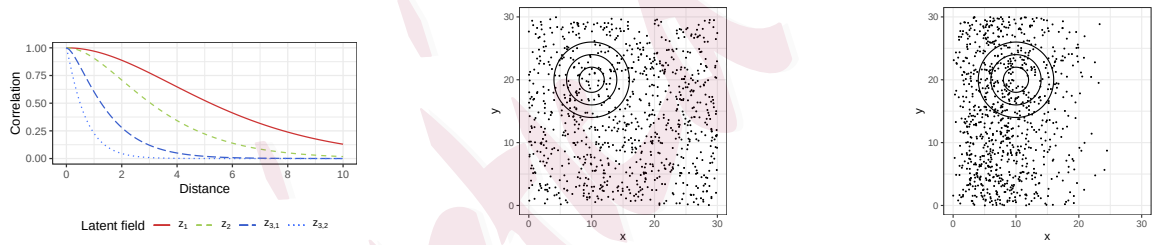

Figure 1: Left: Matérn correlation functions for model setting 1, which consists of the signal random field $(z_1, z_2, z_{3,1})$ with parameters $(\nu, \phi) \in \{(3, 2), (2, 1.5), (1, 1)\}$ and model setting 2 formed by the signal random field $(z_1, z_2, z_{3,2})$ with parameters $(\nu, \phi) \in \{(3, 2), (2, 1.5), (0.6, 0.6)\}$. Middle and right: uniform (middle) 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 the local covariance matrices (2.3), we used two different kernel function settings. Kernel setting 1 used only one ring kernel function (3.7) with parameters $(r_1, r_2) = (0, 2)$, while kernel setting 2 used 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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

circles represent the different ring kernel radii.

For each of the four simulation settings, we carried out 2000 repetitions, and in each repetition we tested three different null hypothesis (H_{02}, H_{03}, and H_{04}) with 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 fixed the number of re-samples to be $B = 200$, and for the full spatial bootstrap, the block size was 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 appeared to maintain the expected rejection rates, which were 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 too small rejection rate 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 performed equally well. Considering the computation time, the latter bootstrap variant might be preferable, as explored in more detail in Section 4.3.

4.2 Simulation Study 2: Hypothesis Testing for Different Signal and Noise Distributions

In this part of the simulations we compare the quality of the introduced tests for data distributions that are non-Gaussian. To do so we keep the exact same simulation outline and the same model settings as in the former section, but we consider a Gaussian and a non-Gaussian distribution for the latent field. The latent field of the Gaussian setting (as in the former section) has a three-variate

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 3: Rejection rates for model setting 1 with Gaussian and non-Gaussian distribution and the uniform sample location pattern based on 2000 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}
	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
30×30	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
	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
40×40	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
	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
50×50	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
	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
60×60	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

signal part and two-variate standard normal noise part. The non-Gaussian 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 non-Gaussian settings have equal second-order spatial dependence but the distributions are different, therefore, differences in the performance of the tests is a result of the different distributions. Moreover, we do not consider the parametric bootstrap tests for these simulations as they are designed for Gaussian distributions.

Tables 3 - 6 summarize the rejection rates based on 2000 simulation repetitions for a signifi-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 4: Rejection rates for model setting 1 with Gaussian and non-Gaussian distribution and the skewed sample location pattern based on 2000 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}
	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
30×30	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
	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
40×40	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
	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
50×50	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
	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
60×60	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

cance level of $\alpha = 0.05$ for model setting 1 and 2 with uniform and skewed coordinate patterns. In the same fashion as the former results these simulations show again the desired rejection rates of 1.00 for H_{02}, 0.05 for H_{03}, and < 0.05 for H_{04} based on $\alpha = 0.05$. The difference between the rejection rates for the Gaussian and non-Gaussian cases are most of the time only in the third decimal place. This means that the tests still keep their good performance even for heavy-tailed non-Gaussian signal and noise distributions, thus, we carry out the subsequent simulations only for the Gaussian case.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 5: Rejection rates for model setting 2 with Gaussian and non-Gaussian distribution and the uniform sample location pattern based on 2000 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}
	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
30×30	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
	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
40×40	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
	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
50×50	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
	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
60×60	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

4.3 Simulation Study 3: Computation Time Comparison

In this simulation, we investigated the computation times for the various test methods. As an illustrative example, we again considered a five-variate latent random field with model setting 1 and bivariate Gaussian noise components. In addition, we kept the same spatial domain sizes, though the sampling sites were changed to be regular defined as $[0, n] \times [0, n] \cap \mathbb{Z}^2$. H_{03} was tested using the five former mentioned test methods with the same number of bootstrap samples and block sizes. The key difference is that each test was carried out with code designed for irregular sample locations as well as code that takes into account simplifications made possible by the fact that the sample locations were regular (e.g., the simplified spatial bootstrap algorithm). Two ring

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 6: Rejection rates for model setting 2 with Gaussian and non-Gaussian distribution and the skewed sample location pattern based on 2000 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}
	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
30×30	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
	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
40×40	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
	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
50×50	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
	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
60×60	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

kernel functions with parameters $(r_1, r_2) \in \{(0,1), (1,\sqrt{2})\}$ were considered for the irregular code, and kernels of the form $f(\mathbf{s}) = I(||\mathbf{s}|| = h)$ with $h \in \{1, \sqrt{2}\}$ were considered for the regular code (one-way and two-way lag-1 local covariance matrices). This choice ensured that the same neighbors were selected for both versions of the code and thus that the qualitative results of the tests were 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 computation times revealed that asymptotic tests are fastest, as the SBSS solution needs to be computed only once, whereas bootstrap algorithms

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

compute the SBSS solution B times.

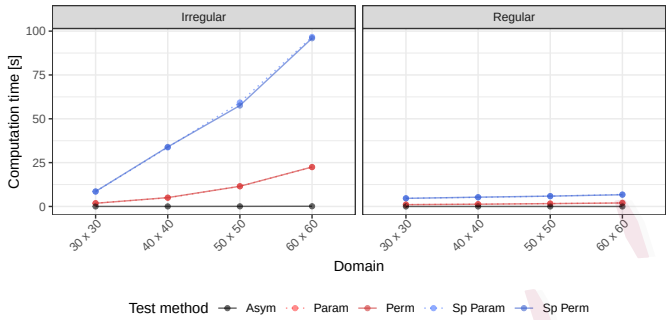

Figure 2: Median running times of the five different test methods for different domain sizes with regular sampling sites based on five simulation repetitions. Computations were carried out with code designed for regular and irregular sampling sites.

Of greater interest is the overall difference in the computation time between regular and irregular code. This might be explained by the fact that the code for regular sampling sites does not rely on distances between sampling sites as the irregular code does. Specifically, the selection of neighbors for local covariance matrices can be implemented 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, as looping through the distance matrix depends on the actual number of locations, while coordinate shifting does not.

Further, there was 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 the impact of the simplified spatial bootstrap variant for regular sampling sites. As explained above, for the irregular code the distance matrix has to be computed again for every new iteration because the spatial bootstrap changes sampling sites for each iteration, which is not the case for the regular code, for which the sampling sites remain equal for each bootstrap

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

iteration.

Overall, this simulation strongly indicates that regular sampling sites should be computationally treated as such. In addition, considering the overall similar performances of the tests in the former simulation, the spatial bootstrapping step for the irregular data might be discarded, as 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 introduced 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) by 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 increasing range parameter. The hypothesis is tested by the asymptotic test and the parametric bootstrap test without full spatial bootstrapping as the performances of all tests in the hypothesis testing simulations were similar but those 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 2000 simulation iterations for uniform and skewed sample location pattern. For lower sample sizes all tests show a desired rejection rate of one at a range parameter of 0.5 which is decreased to 0.3 when the simple size is highest. Interestingly, it seems that there are no differences between the skewed and uniform sample location pattern and the two considered

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

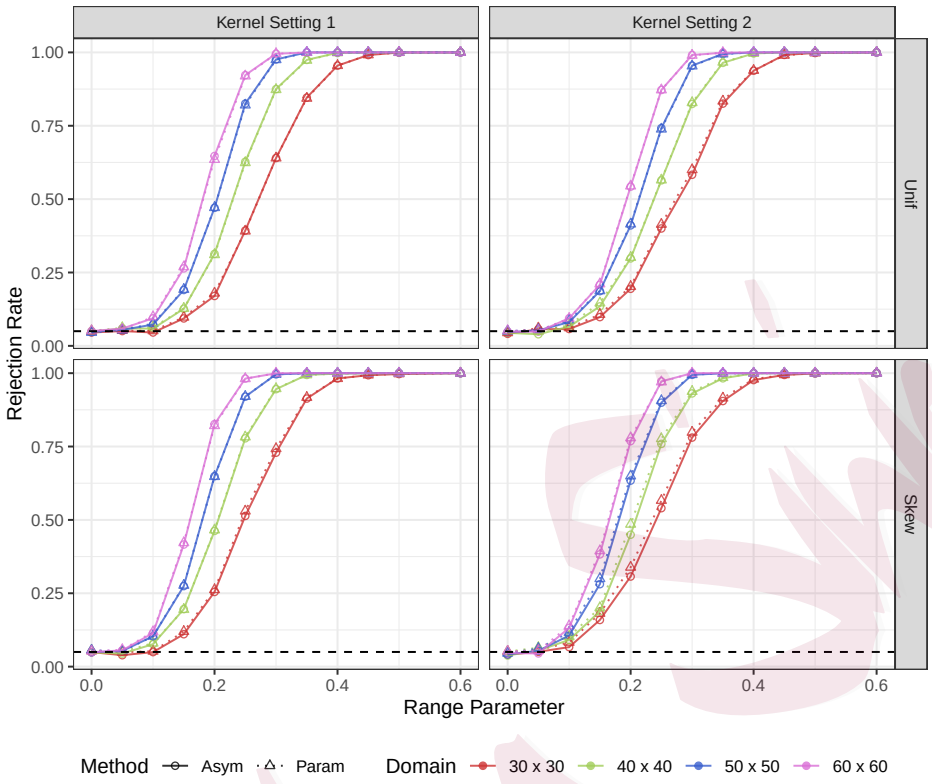

Figure 3: Rejection rates of the asymptotic and parametric bootstrap tests for H_{03} for different kernel settings as a function of the range parameter of the first entry of the signal part at a test significance level of 0.05 (indicated by the dashed line). The results are based on 2000 repetitions.

kernel function settings. Furthermore, this simulation shows no significant difference between the asymptotic and the bootstrap test which again favors the asymptotic one for practical considerations.

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 explore the use of hypothesis tests for signal dimension estimation. We considered the exact same simulation settings as in Section 4.1 but increased the dimension of the noise part to seven leading to a total latent random field dimension of $p = 10$,

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

while the true signal dimension remained $q = 3$. Estimation of the signal dimension was based on the divide-and-conquer strategy described above. As before, all hypothesis tests were carried out using the asymptotic test method and the parametric bootstrap without full spatial bootstrapping. This choice is justified by the similar performance in signal dimension testing of all bootstrap test variants and the fact that the full spatial bootstrap is computationally unfeasible for such a large simulation.

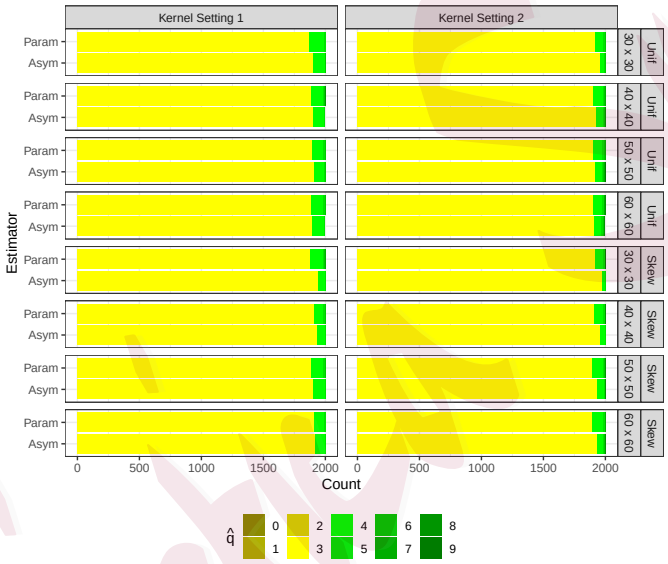

Figure 4: Frequencies of the estimated signal dimension for model setting 1.

Figures 4 and 5 depict the estimated dimensions for 2000 simulation repetitions for a significance level of $\alpha = 0.05$. Overall, the estimation was highly accurate, with the estimated dimension being equal to the true one in approximately 95% of the cases. Interestingly, the signal dimension was never underestimated, while it was overestimated in approximately 100 of the simulation iterations, reflecting the significance level $\alpha = 0.05$. For all settings, the asymptotic test performed better than the bootstrap test. This was especially true for low sample sizes, which is a counter-intuitive result. However, it may be due to the fact that, as the former simulations show, for low sample sizes the asymptotic test never met the theoretical rejection rate, which is simply the sig-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

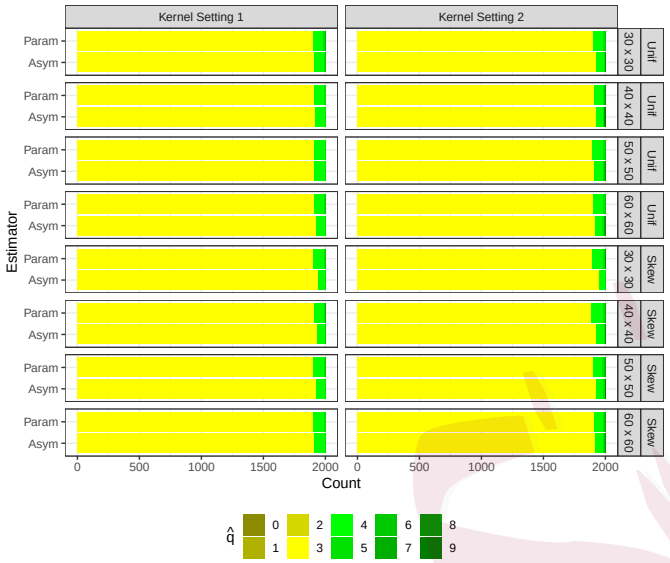

Figure 5: Frequencies of the estimated signal dimension for model setting 2.

nificance level when the null is actually true for small sample sizes (Tables 1 and 2). Therefore, the true null is more often accepted leading to a better performance when estimating the signal dimension.

5. Concluding Remarks

In this paper, we propose and study 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. Besides the theoretical results, the five simulation studies presented in Section 4 demonstrate that our asymptotic tests are comparable to the bootstrap ones in terms of hypothesis testing and estimation. In terms of computation time, our asymptotic method is much faster than the bootstrap ones. When a regular domain structure is used, the computation time can be significantly decreased.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Our proposed dimension tests in the SBSS context might be very 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, leading to a reduction from building a single multivariate model to building several univariate ones. This procedure was already investigated and found to be useful by Muehlmann et al. (2021). As an additional step, one of our proposed dimension tests can be carried out before the spatial prediction, leading to a reduction of the latent field dimension, which results in the need for even fewer univariate models to be built.

However as it is not clear how to obtain in a data-driven way the sequence mentioned in Proposition 4 leading to a consistent estimate we plan in further research to develop a ladle estimator (Luo and Li, 2016, 2021) for this setting which will be based either on bootstrapping or data augmentation. Other future ideas for future research are to develop similar approaches for spatiotemporal data and study the fixed-domain asymptotic properties (Stein, 1995; Cressie, 1993, Section 5.8) of SBSS. Also the SBSS model might be interesting to study in a high-dimension framework, where we could transfer SBSS to a spiked model when both n and p goes to infinity. The problem of identifying the number of spikes has been studied for example in Passemier and Yao (2014). When $p < n$ is diverging, Zhang et al. (2022) proposed recently a new way to estimate the mixing matrix for a SBSS model. So, combining the two methods might provide an insight to select the high-dimensional signal. But we suspect it to be quite hard to investigate the limiting behavior of eigenvalues in a spatial setting.

Supplemental Material

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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Acknowledgments

The author would like to thank the Editor-in-Chief, the 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 are supported by the Austrian Science Fund (No. P31881-N32). The work of Mengxi Yi is also supported by the National Natural Science Foundation of China (No. 12101119).

References

Bachoc, F., J. Betancourt, R. Furrer, and T. Klein (2020). Asymptotic properties of the maximum likelihood and cross validation estimators for transformed Gaussian processes. Electronic Journal of Statistics 14(1), 1962 – 2008.

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

Bevilacqua, M., C. Gaetan, J. Mateu, and E. Porcu (2012). Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. Journal of the American Statistical Association 107(497), 268–280.

Bivand, R. S., E. Pebesma, and V. Gomez-Rubio (2013). Applied spatial data analysis with R, Second edition. Springer, NY.

Bodenham, D. A. and N. M. Adams (2016). A comparison of efficient approximations for a weighted sum of chi-squared random variables. Statistics and Computing 26(4), 917–928.

Bura, E. and R. D. Cook (2001). Extending sliced inverse regression. Journal of the American Statistical Association 96(455), 996–1003.

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

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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Cressie, N. (1993). Statistics for spatial data. John Wiley & Sons.

De Iaco, S., D. Myers, M. Palma, and D. Posa (2013). Using simultaneous diagonalization to identify a space–time linear coregionalization model. Mathematical Geosciences 45(1), 69–86.

Dudley, R. M. (2018). Real analysis and probability. CRC Press.

Emery, X. (2010, September). Iterative algorithms for fitting a linear model of coregionalization. Comput. Geosci. 36(9), 1150–1160.

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

Gneiting, T., W. Kleiber, and M. Schlather (2010). Matérn cross-covariance functions for multivariate random fields. Journal of the American Statistical Association 105(491), 1167–1177.

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

Gräler, B., E. Pebesma, and G. Heuvelink (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. New York: Springer.

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

Luo, W. and B. Li (2016). Combining eigenvalues and variation of eigenvectors for order determination. Biometrika 103(4), 875–887.

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

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

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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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

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

Muehlmann, C., K. Nordhausen, and M. Yi (2021). On cokriging, neural networks, and spatial blind source separation for multivariate spatial prediction. IEEE Geoscience and Remote Sensing Letters 18(11), 1931–1935.

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

Nordhausen, K., H. Oja, P. Filzmoser, and C. Reimann (2015). Blind source separation for spatial compositional data. Mathematical Geosciences 47(7), 753–770.

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

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

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

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

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

Passemier, D. and J. Yao (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. Vienna, Austria: R Foundation for Statistical

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Computing.

Schlather, M., A. Malinowski, P. J. Menck, M. Oesting, and K. Strokorb (2015). Analysis, simulation and prediction of multivariate random fields with package RandomFields. Journal of Statistical Software 63(8), 1–25.

Schmidt, A. M. and A. E. Gelfand (2003). A Bayesian coregionalization approach for multivariate pollutant data. Journal of Geophysical Research: Atmospheres 108(D24).

Shaby, B. and D. Ruppert (2012). Tapered covariance: Bayesian estimation and asymptotics. Journal of Computational and Graphical Statistics 21(2), 433–452.

Stein, M. L. (1995). Fixed-domain asymptotics for spatial periodograms. Journal of the American Statistical Association 90(432), 1277–1288.

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

Von Storch, H. and F. W. Zwiers (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., S. Hao, and Q. Yao (2022). Blind source separation over space.

Computational Statistics, Vienna University of Technology, Austria.
E-mail: christoph.muehlmann@tuwien.ac.at

Institut de Mathématiques de Toulouse, Université Paul Sabatier, France.
E-mail: francois.bachoc@math.univ-toulouse.fr

Computational Statistics, Vienna U. of Technology, Austria; Department of Mathematics and Statistics, U. of Jyväskylä, Finland.
E-mail: klaus.k.nordhausen@jyu.fi

School of Statistics, Beijing Normal University, China; Computational Statistics, Vienna University of Technology, Austria.
E-mail: mxyi@bnu.edu.cn