Biometrika (2020), **107**, 3, *pp.* 627–646
Printed in Great Britain
doi: 10.1093/biomet/asz079
Advance Access publication 17 February 2020

Spatial blind source separation

BY FRANÇOIS BACHOC

Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse, France

francois.bachoc@math.univ-toulouse.fr

MARC G. GENTON

Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

marc.genton@kaust.edu.sa

KLAUS NORDHAUSEN

Computational Statistics, Vienna University of Technology, Wiedner Hauptstr. 7, A-1040 Vienna, Austria

klaus.nordhausen@tuwien.ac.at

ANNE RUIZ-GAZEN

Toulouse School of Economics, University of Toulouse Capitole, 1, Esplanade de l'Université, 31080 Toulouse Cedex 06, France

anne.ruiz-gazen@tse-fr.eu

AND JONI VIRTA

Department of Mathematics and Statistics, University of Turku, 20014 Turun yliopisto, Finland

joni.virta@utu.fi

SUMMARY

Recently a blind source separation model was suggested for spatial data, along with an estimator based on the simultaneous diagonalization of two scatter matrices. The asymptotic properties of this estimator are derived here, and a new estimator based on the joint diagonalization of more than two scatter matrices is proposed. The asymptotic properties and merits of the novel estimator are verified in simulation studies. A real-data example illustrates application of the method.

Some key words: Joint diagonalization; Limiting distribution; Multivariate random field; Spatial scatter matrix.

1. INTRODUCTION

Multivariate data measured at spatial locations $s_1, \ldots, s_n$ in a domain $\mathcal{S}^d \subseteq \mathbb{R}^d$ are frequently encountered. Such data exhibit two kinds of dependence: measurements taken closer to each other tend to be more similar than measurements taken further apart, and the variable values within a single location are likely to be correlated.

This complexity makes modelling multivariate spatial data computationally and theoretically difficult due to the large number of parameters required to represent the dependencies. In the present article we address this problem through blind source separation, a framework established for independent component analysis of independent and identically distributed data and of stationary and nonstationary time series; see Comon & Jutten (2010) and Nordhausen & Oja (2018). Denoting a p-variate random field by $X(s) = \{X_1(s), \ldots, X_p(s)\}^{\mathrm{T}}$, where $^{\mathrm{T}}$ is the transpose operator, we assume that $X(s)$ obeys the spatial blind source separation model introduced in Nordhausen et al. (2015); that is, $X(s)$ at location s is a linear mixture of an underlying p-variate latent field $Z(s) = \{Z_1(s), \ldots, Z_p(s)\}^{\mathrm{T}}$ with independent components,

$$X(s) = \Omega Z(s), \tag{1}$$

where Ω is an unknown $p \times p$ full-rank matrix. In this introductory section, we treat the random fields X and Z as having mean functions zero for simplicity.

When the observed random field X takes the form (1), modelling and computational simplifications can be obtained. In fact, if no assumption at all is made on X, then the distribution of X is characterized by p covariance functions and $p(p-1)/2$ cross-covariance functions. In contrast, when it is assumed that X takes the form (1), then the distribution of X is characterized by p covariance functions and a $p \times p$ matrix. A function, being an infinite-dimensional object, is more difficult to model and estimate than a fixed-dimensional matrix. Therefore, when the observed random field X takes the form (1), modelling simplifications are available.

When no assumption is made on X, a common practice in geostatistics is to let each of the p covariance functions and each of the $p(p-1)/2$ cross-covariance functions of X be characterized by q parameters. For example, the case of $q = 2$ could correspond to a variance and a length-scale parameter for an isotropic function. Then the resulting $qp(p+1)/2$ parameters are usually estimated jointly by optimizing a fit criterion, typically the likelihood (Genton & Kleiber, 2015). This involves solving an optimization problem in dimension $qp(p+1)/2$, where the computational cost of an evaluation of the likelihood is $O(p^3n^3)$. Once the $qp(p+1)/2$ parameters are estimated, the prediction of $X(s)$ for new values of s can be performed at a computational cost of $O(p^3n^3)$.

In contrast, suppose that model (1) holds for X. We will show in this paper that an estimate of Ω^{-1} can be obtained. This is done by first computing scatter matrices with computational cost $O(p^2n^2)$ and then performing an optimization in dimension p^2, where the computational cost of the function to be evaluated is $O(p^2)$; see § 4 for details. If each covariance function of Z is characterized by q parameters, then each can be estimated separately by optimizing the likelihood in dimension q. The evaluation cost of the likelihood is $O(n^3)$. Once the qp covariance parameters are estimated, the prediction of $X(s)$ for new values of s can be performed at $O(pn^3)$ cost. Indeed, the predictions of $Z_1(s), \ldots, Z_p(s)$ can be performed separately at $O(n^3)$ cost and then aggregated at negligible cost.

Not all random fields X obey a spatial blind source separation model of the form (1). For instance, (1) forces the cross-covariance functions of X to be symmetric. Nevertheless, it is a reasonable model in a fair number of practical situations (Nordhausen et al., 2015) and brings the computational benefits discussed above. Furthermore, an additional benefit of the form (1) is dimension reduction. In blind source separation, often significantly fewer than the full p latent components are needed to capture the essential structure of the original observations, and the remaining components can be discarded as noise.

We therefore consider the spatial blind source separation model (1) in this paper and focus on the estimation of Ω^{-1}. As discussed above, this estimation enables us to estimate the cross-covariance functions of X and to perform prediction. Our approach to estimating Ω^{-1} is based on the use of local covariance, or scatter, matrices,

$$\hat{M}(f) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} f(s_i - s_j) X(s_i) X(s_j)^{\mathrm{T}}, \tag{2}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is called the kernel function. Nordhausen et al. (2015) obtained estimators $\hat{\Gamma}(f)$ of Ω^{-1} through a generalized eigendecomposition of pairs of local covariance matrices with kernels of the form (f_0, f_h), where $f_h(s_i - s_j) = I(\|s_i - s_j\| \leqslant h)$ for a positive constant h and $f_0(s) = I(s = 0)$, with $I(\cdot)$ denoting the indicator function. The estimators of Nordhausen et al. (2015) are based on the following definition, with $f = f_h$ for some $h > 0$.

DEFINITION 1. *An unmixing matrix estimator $\hat{\Gamma}(f)$ jointly diagonalizes $\hat{M}(f_0)$ and $\hat{M}(f)$ in the following way:*

$$\hat{\Gamma}(f)\hat{M}(f_0)\hat{\Gamma}(f)^{\mathrm{T}} = I_p, \quad \hat{\Gamma}(f)\hat{M}(f)\hat{\Gamma}(f)^{\mathrm{T}} = \hat{\Lambda}(f),$$

where $\hat{\Lambda}(f)$ is a diagonal matrix with diagonal elements arranged in decreasing order.

This method is conceptually close to principal component analysis, where latent variables that have maximal variance are found through diagonalization of the covariance matrix. However, since the covariance matrix does not capture spatial information, it was extended to the concept of local covariance matrix in Nordhausen et al. (2015). Analogously, the diagonalization of local covariance matrices aims to find latent fields that maximize spatial correlation.

Here, we expand on the work of Nordhausen et al. (2015) by relaxing the condition on the kernel f in Definition 1 so that it is no longer restricted to be of the ball form f_h. Furthermore, we derive the asymptotic behaviour of the method proposed in Nordhausen et al. (2015) for a large class of kernel functions f.

The idea behind the construction of these kernel functions is that the mean values of $\hat{M}(f)$ and $\hat{M}(f_0)$ would be diagonal matrices if, in their definition, the mixed components X were replaced by the latent components Z. Hence, a general blind source separation strategy is to undo the mixing in X by finding a matrix $\hat{\Gamma}(f)$ that simultaneously diagonalizes $\hat{M}(f)$ and $\hat{M}(f_0)$. This task is computationally simple and can always be done exactly using generalized eigenvalue-eigenvector theory. However, from temporal blind source separation, it is well known that when diagonalizing only two matrices, the choice of the matrices can have a large impact on the separation efficiency. Therefore, a popular strategy is to approximately diagonalize more than two matrices in the hope of including more information; see, for example, Belouchrani et al. (1997), Miettinen et al. (2014), Nordhausen (2014), Matilainen et al. (2015) and Miettinen et al. (2016). Approximate diagonalization becomes necessary as the matrices commute only at the population level, but not when estimated using finite data. There are many algorithms available for this purpose. We use this idea to extend the method of Nordhausen et al. (2015) to the joint diagonalization of more than two local covariance matrices. We also derive the asymptotic behaviour of the proposed estimators.

2. SPATIAL BLIND SOURCE SEPARATION MODEL

2.1. *General assumptions*

In the spatial blind source separation model, the following assumptions are made.

Assumption 1. We have $E\{Z(s)\} = 0$ for $s \in \mathcal{S}^d$.

Assumption 2. We have $\mathrm{cov}\{Z(s)\} = E\{Z(s)Z(s)^{\mathrm{T}}\} = I_p$.

Assumption 3. We have $\text{cov}\{Z(s_1), Z(s_2)\} = E\{Z(s_1)Z(s_2)^{\mathrm{T}}\} = D(s_1, s_2)$, where D is a diagonal matrix whose diagonal elements depend only on $s_1 - s_2$.

Let $\text{cov}\{Z_k(s_i), Z_k(s_j)\} = K_k(s_i - s_j) = D(s_i, s_j)_{k,k}$, where K_k denotes the stationary covariance function of Z_k for $k = 1, \ldots, p$.

Assumption 1 is made for convenience and can easily be replaced by the assumption of a constant unknown mean, as shown in the Supplementary Material. Assumption 2 requires that the components of $Z(s)$ be uncorrelated and implies that the variances of the components are equal to 1, which alleviates identifiability issues and holds without loss of generality. Assumption 3 says that there is also no spatial cross-dependence between the components. However, even after these assumptions are made, the model is not uniquely defined. The order of the latent fields and also their signs can be changed. This is common to all blind source separation approaches and is not found to be a problem in practice.

2.2. *Identifiability*

The expectations of $\hat{M}(f)$ and $\hat{M}(f_0)$ are, respectively,

$$M(f) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} f(s_i - s_j) E\{X(s_i)X(s_j)^{\mathrm{T}}\}, \qquad M(f_0) = n^{-1} \sum_{i=1}^{n} E\{X(s_i)X(s_i)^{\mathrm{T}}\}.$$

Hence the empirical procedure of Definition 1, operating on $\hat{M}(f)$ and $\hat{M}(f_0)$, can be associated with the following theoretical procedure operating on $M(f)$ and $M(f_0)$.

DEFINITION 2. *For any function $f : \mathbb{R}^d \to \mathbb{R}$, an unmixing matrix functional $\Gamma(f)$ is defined to be a functional which jointly diagonalizes $M(f)$ and $M(f_0)$ in the following way:*

$$\Gamma(f)M(f_0)\Gamma(f)^{\mathrm{T}} = I_p, \qquad \Gamma(f)M(f)\Gamma(f)^{\mathrm{T}} = \Lambda(f),$$

where $\Lambda(f)$ is a diagonal matrix with diagonal elements arranged in decreasing order.

We remark that an unmixing matrix $\Gamma(f)$ can be found using the generalized eigenvalue-eigenvector theory. In addition, an unmixing matrix is never unique, since if $\Gamma(f)$ and $\Lambda(f)$ satisfy Definition 2, then $S\Gamma(f)$ and $\Lambda(f)$ also satisfy Definition 2 for any diagonal matrix S with diagonal elements equal to -1 or 1. We also remark that $\Lambda(f)$ is not the expectation of $\hat{\Lambda}(f)$ in general. Indeed, Definitions 1 and 2 are based on nonlinear functions of $\{\hat{M}(f), \hat{M}(f_0)\}$ and of $\{M(f), M(f_0)\}$, respectively.

The usual notion of identifiability in blind source separation is that any unmixing functional $\Gamma(f)$ should recover the components of Z up to signs and order of the components. Thus, any unmixing functional $\Gamma(f)$ should coincide with Ω^{-1} up to the order and signs of the rows.

DEFINITION 3. *We say that the unmixing problem given by f is identifiable if any unmixing functional $\Gamma(f)$ satisfying Definition 2 can be written as $PS\,\Omega^{-1}$, where P is a permutation matrix and S is a diagonal matrix with diagonal elements equal to -1 or 1.*

The motivation behind identifiability is that if identifiability holds, then estimating $M(f_0)$ and $M(f)$ consistently by $\hat{M}(f_0)$ and $\hat{M}(f)$ enables us to obtain $\hat{\Gamma}(f)$, which will be approximately equal to a matrix of the form $PS\,\Omega^{-1}$, with P and S as in Definition 3. The following proposition provides a necessary and sufficient condition for identifiability. This proposition and all other

theoretical results in the paper are proved in the Supplementary Material. Let $M^{-\mathrm{T}}$ denote the inverse of the transpose of M.

PROPOSITION 1. *The unmixing problem given by f is identifiable if and only if the diagonal elements of $\Omega^{-1}M(f)\Omega^{-\mathrm{T}}$ are distinct.*

We remark that identifiability is a joint property of the kernel f and the covariance functions $K_1, \ldots, K_p$. For example, consider the situation in which $K_1, \ldots, K_p$ are compactly supported and equal to zero at distances larger than $0 < r < \infty$, where the function $f(s) = I(r_1 < \|s\| \leqslant r_2)$ with $r \leqslant r_1 < r_2 < \infty$ is used as the kernel. Then identifiability does not hold because $\Omega^{-1}M(f)\Omega^{-\mathrm{T}}$ is equal to the zero matrix. On the other hand, if f is a ball kernel of the form $f(s) = I(\|s\| \leqslant r_0)$ with $r_0 > 0$, then identifiability may hold for the same covariance functions $K_1, \ldots, K_p$.

Finally, for any kernel f, a necessary condition for identifiability is that there should not exist $k, l \in \{1, \ldots, p\}$ with $k \neq l$ such that $K_k(s_i - s_j) = K_l(s_i - s_j)$ for all $i, j = 1, \ldots, n$. Indeed, if this were the case, then the diagonal elements k and l of $\Omega^{-1}M(f)\Omega^{-\mathrm{T}}$ would be equal for any kernel f. An extreme example of this situation is where $K_1 = \cdots = K_p$ with only Gaussian components. In this case, for any orthogonal matrix Q, the distribution of the random field QZ is the same as that of the random field Z; hence no statistical procedure can be expected to recover the components of Z, even up to signs and permutations, when one only observes the transformed random field X.

2.3. *Relationships to other models of multivariate random fields*

The spatial blind source separation model is notably different from the usual multivariate models for spatial data, which are often defined starting with their covariance functions contained in a cross-covariance matrix,

$$C(s_1, s_2) = \mathrm{cov}\{X(s_1), X(s_2)\} = \{C_{k,l}(s_1, s_2)\}_{k,l=1}^{p},$$

whereas our method for estimating Ω^{-1} does not need to model or estimate the covariance functions of the latent fields $Z_1(s), \ldots, Z_p(s)$.

In a recent extensive review, Genton & Kleiber (2015) discussed different approaches to defining cross-covariance matrix functionals and gave a list of properties and conventions that they should satisfy, including stationarity and invariance under rotation. As Genton & Kleiber (2015) pointed out, to create general classes of models with well-defined cross-covariance functionals is a major challenge. Multivariate spatial models are particularly challenging as many parameters need to be fitted. In textbooks such as Wackernagel (2003), usually the following two popular models are described.

In the intrinsic correlation model it is assumed that the stationary covariance matrix $C(h)$ can be written as the product of the variable covariances and the spatial correlations, $C(h) = \rho(h)T$ for all lags h, where T is a nonnegative-definite $p \times p$ matrix and $\rho(h)$ a univariate spatial correlation function.

The more popular linear model of coregionalization is a generalization of the intrinsic correlation model, and the covariance matrix has the form

$$C(h) = \sum_{k=1}^{r} \rho_k(h) T_k$$

for some positive integer $r \leqslant p$, where all the ρ_k are univariate spatial correlation functions and the T_k are nonnegative-definite $p \times p$ matrices, often called coregionalization matrices. With $r = 1$ this reduces to the intrinsic correlation model. The linear model of coregionalization implies a symmetric cross-covariance matrix.

Estimation in the linear model of coregionalization is discussed in several papers. Goulard & Voltz (1992) studied the coregionalization matrices using an iterative algorithm where the spatial correlation functions are assumed to be known. The algorithm was extended by Emery (2010). Assuming Gaussian random fields, an expectation-maximization algorithm was suggested in Zhang (2007), and a Bayesian approach was considered in Gelfand et al. (2004).

There is a simple connection between the spatial blind source separation model and the linear model of coregionalization. The covariance matrix $C_X(h)$ resulting from a spatial blind source separation model is always symmetric and can be written as

$$C_X(h) = \sum_{k=1}^{p} K_k(h) T_k,$$

where $T_k = \omega_k \omega_k^{\mathrm{T}}$, with ω_k being the kth column of Ω. Thus, the spatial blind source separation model is a special case of the linear model of coregionalization with $r = p$, and with all the coregionalization matrices T_k $(k = 1, \ldots, p)$ being rank-one matrices.

3. Asymptotic properties for simultaneous diagonalization of two matrices

Recall the definition (2) of a local covariance matrix and that

$$\hat{M}(f_0) = n^{-1} \sum_{i=1}^{n} X(s_i) X(s_i)^{\mathrm{T}}$$

is the covariance estimator. Asymptotic results can be derived for the previous estimators under Assumptions 1–3 and the further assumptions below.

Assumption 4. The coordinates $Z_1, \ldots, Z_p$ of Z are stationary Gaussian processes on $\mathbb{R}^d$.

Assumption 5. A fixed $\Delta > 0$ exists such that for all $n \in \mathbb{N}$ and all $i, j = 1, \ldots, n$ with $i \neq j$, $\|s_i - s_j\| \geqslant \Delta$.

Assumption 6. Fixed $A > 0$ and $\alpha > 0$ exist such that for all $x \in \mathbb{R}^d$ and all $k = 1, \ldots, p$,

$$|K_k(x)| \leqslant \frac{A}{1 + \|x\|^{d+\alpha}}.$$

Assumption 7. If Assumption 6 holds, then for the same $A > 0$ and $\alpha > 0$ we have

$$|f(x)| \leqslant \frac{A}{1 + \|x\|^{d+\alpha}}.$$

Assumption 8. We have

$$\liminf_{n \to \infty} \min_{i=2,\ldots,p} \left[\left\{ \Omega^{-1} M(f) \Omega^{-\mathrm{T}} \right\}_{i,i} - \left\{ \Omega^{-1} M(f) \Omega^{-\mathrm{T}} \right\}_{i-1,i-1} \right] > 0.$$

Assumption 5 implies that $\mathcal{S}^d$ is unbounded as $n \to \infty$, which means that we are dealing with the increasing-domain asymptotic framework (Cressie, 1993).

Assumption 7 holds in particular for the function $I(s = 0)$ and for the ball and ring kernels $B(h)(s) = I(\|s\| \leqslant h)$ with fixed $h \geqslant 0$ and $R(h_1, h_2)(s) = I(h_1 \leqslant \|s\| \leqslant h_2)$ with fixed $h_2 \geqslant h_1 \geqslant 0$.

Up to reordering of the components of Z, which can be done without loss of generality, Assumption 8 is an asymptotic version of the identifiability condition in Proposition 1. Under Assumption 8, identifiability in the sense of Definition 3 holds for sufficiently large n, by Proposition 1.

Proposition 2 below establishes the consistency of the estimator $\hat{M}(f)$, where f satisfies Assumption 7.

PROPOSITION 2. *Suppose $n \to \infty$ and that Assumptions 1–6 hold, and let $f : \mathbb{R}^d \to \mathbb{R}$ satisfy Assumption 7. Then $\hat{M}(f) - M(f) \to 0$ in probability as $n \to \infty$.*

We remark that $M(f)$ depends on n and that we do not assume that the sequence of matrices $M(f)$ converges to a fixed matrix as $n \to \infty$. Hence, Proposition 2 shows that $\hat{M}(f) - M(f)$ converges to zero, not that $\hat{M}(f)$ converges to $M(f)$.

Next, we show the joint asymptotic normality of $n^{1/2}\{\hat{M}(f_0) - M(f_0)\}$ and $n^{1/2}\{\hat{M}(f) - M(f)\}$, viewed as sequences of $p^2 \times 1$ random vectors. As in Proposition 2, we do not need to assume that the sequence of $2p^2 \times 2p^2$ covariance matrices of these two sequences of vectors converges to a fixed matrix. Hence, we will not show that these sequences of random vectors converge jointly to a fixed Gaussian distribution; instead, we show that the distances between the distributions of these random vectors and Gaussian distributions converge to zero as $n \to \infty$. As a distance between distributions, we consider a metric d_{w} generating the topology of weak convergence on the set of Borel probability measures on Euclidean spaces (see, e.g., Dudley, 2002, p. 393). The advantage of using such a distance is that a sequence of distributions $(\mathcal{L}_n)_{n\in\mathbb{N}}$ converges to a fixed distribution $\mathcal{L}$ if and only if $d_{\mathrm{w}}(\mathcal{L}_n, \mathcal{L})$ converges to zero. The next proposition provides the asymptotic normality result.

PROPOSITION 3. *Under the same assumptions as in Proposition 2, let $W(f)$ be the vector of size $p^2 \times 1$ defined, for $i = (a-1)p + b$ with $a, b \in \{1, \ldots, p\}$, by*

$$W(f)_i = n^{1/2}\{\hat{M}(f)_{a,b} - M(f)_{a,b}\}.$$

Let Q_n be the distribution of $\{W(f)^{\mathrm{T}}, W(f_0)^{\mathrm{T}}\}^{\mathrm{T}}$. Then, as $n \to \infty$,

$$d_{\mathrm{w}}\big[Q_n, N\{0, V(f, f_0)\}\big] \to 0,$$

where N denotes the normal distribution and details of the matrix $V(f, f_0)$ are given in the Appendix. Moreover, the largest eigenvalue of $V(f, f_0)$ is bounded as $n \to \infty$.

In Proposition 3, $V(f, f_0)$ is a $2p^2 \times 2p^2$ matrix that depends on n and is interpreted as an asymptotic covariance matrix. Also, in Proposition 3 the vectors $W(f)$ and $W(f_0)$, which are asymptotically Gaussian, are obtained by row vectorization of $n^{1/2}\{\hat{M}(f_0) - M(f_0)\}$ and $n^{1/2}\{\hat{M}(f) - M(f)\}$. Taking $f(s) = I(\|s\| \leqslant h)$ with $h > 0$ in Propositions 2 and 3 gives the asymptotic properties of the method proposed in Nordhausen et al. (2015).

Remark 1. Propositions 2 and 3 remain valid when the process X is centred by $\bar{X} = n^{-1}\sum_{i=1}^{n} X(s_i)$. Indeed, we prove in the Supplementary Material that the difference between the centred estimator and $\hat{M}(f)$ is of order $O_{\mathrm{p}}(n^{-1})$.

For a matrix A with rows $l_1^{\mathrm{T}}, \ldots, l_k^{\mathrm{T}}$, let $\mathrm{vect}(A) = (l_1^{\mathrm{T}}, \ldots, l_k^{\mathrm{T}})^{\mathrm{T}}$ be the row vectorization of A; for a matrix A of size $k \times k$, let $\mathrm{diag}(A) = (A_{1,1}, \ldots, A_{k,k})^{\mathrm{T}}$. The next proposition shows the joint asymptotic normality of the estimators $\hat{\Gamma}(f)$ and $\hat{\Lambda}(f)$.

PROPOSITION 4. *Under the same assumptions as in Proposition 2, suppose also that Assumption 8 holds. For $\hat{\Gamma}(f)$ and $\hat{\Lambda}(f)$ in Definition 1, let Q_n be the distribution of*

$$n^{1/2} \begin{bmatrix} \mathrm{vect}\{\hat{\Gamma}(f) - \Omega^{-1}\} \\ \mathrm{diag}\{\hat{\Lambda}(f) - \Lambda(f)\} \end{bmatrix}.$$

Then we can choose $\hat{\Gamma}(f)$ and $\hat{\Lambda}(f)$ in Definition 1 such that as $n \to \infty$,

$$d_{\mathrm{w}}\{Q_n, N(0, F_1)\} \to 0,$$

where the matrix F_1 is detailed in the Appendix.

In Proposition 4, as before we consider the sequences of vectors obtained by vectorizing $n^{1/2}\{\hat{\Gamma}(f) - \Omega^{-1}\}$ and taking the diagonal of $n^{1/2}\{\hat{\Lambda}(f) - \Lambda(f)\}$. Again, we do not show that the sequence of joint distributions of these vectors converges to a fixed distribution, but rather show that these joint distributions are asymptotically close to Gaussian distributions, with covariance matrices given by F_1. We remark that F_1 denotes a sequence of $(p^2+p) \times (p^2+p)$ matrices. We also remark that in Definition 1, $\hat{\Gamma}(f)$ is not uniquely defined; it is defined up to the signs of its rows. Hence, Proposition 4 shows that there exists a choice of the sequence $\hat{\Gamma}(f)$ in Definition 1 such that asymptotic normality holds as $n \to \infty$.

The performance of the estimators $\hat{\Gamma}(f)$ and $\hat{\Lambda}(f)$ depends on the choice of $\hat{M}(f)$, which should be chosen so that $\hat{\Lambda}(f)$ has diagonal elements as distinct as possible. This is similar to the time series context as described in Miettinen et al. (2012). To avoid this dependency in the time series context, the joint diagonalization of more than two matrices has been suggested, and we will apply this same idea to the spatial context in the following section.

4. IMPROVING ESTIMATION BY JOINTLY DIAGONALIZING MORE THAN TWO MATRICES

Spatial blind source separation with more than two kernel functions of the form $f_0, f_1, \ldots, f_k$, with $k \geqslant 2$, can be formulated as

$$\hat{\Gamma} \in \underset{\substack{\Gamma:\, \Gamma \hat{M}(f_0) \Gamma^{\mathrm{T}} = I_p, \\ \Gamma \text{ has rows } \gamma_1^{\mathrm{T}}, \ldots, \gamma_p^{\mathrm{T}}}}{\arg\max} \sum_{l=1}^{k} \sum_{j=1}^{p} \{\gamma_j^{\mathrm{T}} \hat{M}(f_l) \gamma_j\}^2. \tag{3}$$

We can show that if $k = 1$, the set of $\hat{\Gamma}$ satisfying (3) coincides with the set of $\hat{\Gamma}(f_1)$ satisfying Definition 1. From experience in the time series blind source separation context (see, e.g., Miettinen et al., 2016), usually the diagonalization of several matrices gives better separation than diagonalization based on two matrices only. In this paper we show that using $k \geqslant 2$ is indeed beneficial both from a theoretical point of view and in practice.

The identifiability notion in Definition 3 and Proposition 1 can be extended to the case of more than two local covariance matrices. We first remark that the theoretical version of (3) is

$$\Gamma \in \underset{\substack{\Gamma:\,\Gamma M(f_0)\Gamma^{\mathrm{T}}=I_p,\\ \Gamma \text{ has rows } \gamma_1^{\mathrm{T}},\ldots,\gamma_p^{\mathrm{T}}}}{\arg\max} \sum_{l=1}^{k}\sum_{j=1}^{p}\{\gamma_j^{\mathrm{T}} M(f_l)\gamma_j\}^2. \tag{4}$$

We then extend Definition 3 and Proposition 1 to the case of more than two local covariance matrices.

DEFINITION 4. *We say that the unmixing problem given by $f_1,\ldots,f_k$ is identifiable if any unmixing functional Γ satisfying (4) can be written as $PS\,\Omega^{-1}$, where P is a permutation matrix and S is a diagonal matrix with diagonal elements equal to -1 or 1.*

PROPOSITION 5. *The unmixing problem given by $f_1,\ldots,f_k$ is identifiable if and only if for every pair $i,j = 1,\ldots,p$ with $i \neq j$, there exists $l = 1,\ldots,k$ such that $\{\Omega^{-1}M(f_l)\Omega^{-\mathrm{T}}\}_{i,i} \neq \{\Omega^{-1}M(f_l)\Omega^{-\mathrm{T}}\}_{j,j}$.*

We remark that the identifiability condition in Proposition 5 is weaker than that in Proposition 1, because if the condition in Proposition 1 holds with f being one of the $f_1,\ldots,f_k$, then the condition in Proposition 5 holds. This is one of the benefits of jointly diagonalizing more than two matrices.

One of the main theoretical contributions of this paper is to provide an asymptotic analysis of the joint diagonalization of several matrices in the spatial context. Assumption 8, on asymptotic identifiability, can be replaced by the following weaker assumption.

Assumption 9. A fixed $\delta > 0$ and an $n_0 \in \mathbb{N}$ exist such that for all $n \in \mathbb{N}$ with $n \geqslant n_0$ and for every pair $i,j = 1,\ldots,p$ with $i \neq j$, there exists $l = 1,\ldots,k$ such that $|\{\Omega^{-1}M(f_l)\Omega^{-\mathrm{T}}\}_{i,i} - \{\Omega^{-1}M(f_l)\Omega^{-\mathrm{T}}\}_{j,j}| \geqslant \delta$.

In the next proposition we state the consistency of $\hat{\Gamma}$.

PROPOSITION 6. *Suppose that Assumptions 1–6 hold. Let $k \in \mathbb{N}$ be fixed, and let $f_1,\ldots,f_k : \mathbb{R}^d \to \mathbb{R}$ satisfy Assumption 7. Further, suppose that Assumption 9 holds. Let $\hat{\Gamma} = \hat{\Gamma}\{\hat{M}(f_0),\hat{M}(f_1),\ldots,\hat{M}(f_k)\}$ satisfy (3). Then we can choose $\hat{\Gamma}$ so that $\hat{\Gamma} \to \Omega^{-1}$ in probability as $n \to \infty$.*

We remark that in Proposition 6, $\hat{\Gamma}$ is defined only up to permutation of the rows and multiplication of the rows by 1 or -1. Hence, we show that there exists a choice of a sequence $\hat{\Gamma}$ that converges to Ω^{-1}. The next proposition provides an asymptotic normality result.

PROPOSITION 7. *Under the same assumptions as in Proposition 6, let $(\hat{\Gamma}_n)_{n\in\mathbb{N}}$ be any sequence of $p \times p$ matrices such that for any $n \in \mathbb{N}$, $\hat{\Gamma}_n = \hat{\Gamma}_n\{\hat{M}(f_0),\hat{M}(f_1),\ldots,\hat{M}(f_k)\}$ satisfies (3). Then there exist a sequence of permutation matrices (P_n) and a sequence of diagonal matrices (D_n) with diagonal elements in $\{-1,1\}$ such that the distribution Q_n of $n^{1/2}\,\mathrm{vect}(\check{\Gamma}_n - \Omega^{-1})$ with $\check{\Gamma}_n = D_nP_n\hat{\Gamma}_n$ satisfies $d_{\mathrm{w}}\{Q_n, N(0,F_k)\} \to 0$ as $n \to \infty$, where the matrix F_k is detailed in the Appendix.*

In Proposition 7, for any $n \in \mathbb{N}$, the choice of $\hat{\Gamma}_n$ satisfying (3) is not unique. The proposition shows that for any choice of the sequence of matrices $\hat{\Gamma}_n$, one can exchange the rows and multiply them by 1 or -1 to obtain a sequence of matrices $\check{\Gamma}_n$ that converges to Ω^{-1} as $n \to \infty$. Furthermore, as in Proposition 4, we show that the sequence of distributions of $n^{1/2}\,\mathrm{vect}(\check{\Gamma}_n - \Omega^{-1})$ is asymptotically close to a sequence of Gaussian distributions. The sequence of $p^2 \times p^2$ covariance matrices of these Gaussian distributions is F_k.

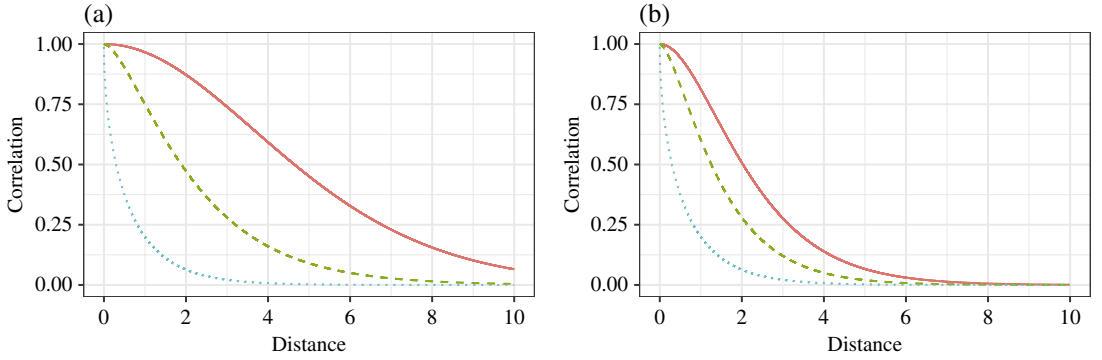

Fig. 1. Matérn covariance functions of the first (solid red line), second (dashed green line) and third (dotted blue line) latent fields used in (a) the simulation of § 5.2 and (b) the simulation of § 5.3. The parameter vectors (κ, ϕ) of the three fields are taken to be $(6, 1.2)$, $(1, 1.5)$ and $(0.25, 1)$ in (a) and $(2, 1)$, $(1, 1)$ and $(0.25, 1)$ in (b).

The idea of joint diagonalization is not new in spatial data analysis. For example, in a model-free context, matrix variograms have been jointly diagonalized in Xie & Myers (1995), Xie et al. (1995) and De Iaco et al. (2013). However, the unmixing matrix was restricted to be orthogonal, which would therefore not solve the spatial blind source separation model.

While two symmetric matrices can always be simultaneously diagonalized, this is usually not the case for more than two matrices that are estimated based on finite data. Therefore, algorithms are needed for approximate joint diagonalization. In this paper we use an algorithm based on Givens rotations (Clarkson, 1988). Other possible algorithms and their effects on the properties of the estimates are discussed in Illner et al. (2015), for example.

5. SIMULATIONS

5.1. *Preliminaries*

In this section we use simulated data to verify our asymptotic results and to compare the efficiencies of the different local covariance estimates under various spatial models. All simulations are performed in `R` (R Development Core Team, 2020) using the packages `geoR` (Ribeiro Jr & Diggle, 2016), `JADE` (Miettinen et al., 2017) and `RcppArmadillo` (Eddelbuettel & Sanderson, 2014). To generate the simulation data, we choose particular covariance functions for the latent fields. However, our proposed methods do not use this information in any way, but operate solely through the selection of local covariance matrices.

5.2. *Asymptotic approximation of the unmixing matrix estimator*

We start with a simple simulation to establish the validity of the asymptotic approximation of the unmixing matrix estimator $\hat{\Gamma}(f)$ for different kernels f and to obtain some preliminary comparative results on the proposed estimators. We consider a centred three-variate spatial blind source separation model $X(s) = \Omega Z(s)$, where each of the three independent latent fields has a Matérn covariance function with shape and range parameters $(\kappa, \phi) \in \{(6, 1.2), (1, 1.5), (0.25, 1)\}$; see Fig. 1(a). We recall that the Matérn correlation function is defined by

$$\rho(h) = 2^{1-\kappa}\,\Gamma(\kappa)^{-1}\,(h/\phi)^{\kappa}\,K_{\kappa}(h/\phi),$$

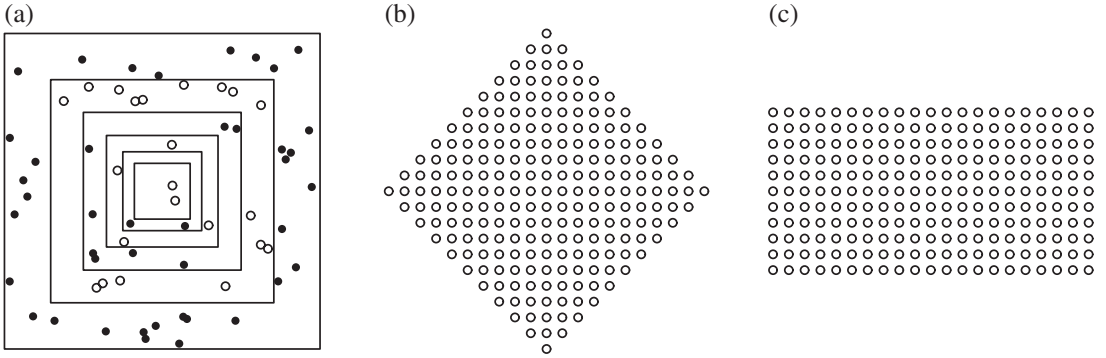

Fig. 2. (a) The location pattern scheme used in § 5.2 with the marker type alternating between the consecutive layers and only 1% of the locations shown for clarity; (b) a diamond grid of radius 10 having $n = 221$ locations; and (c) a rectangular grid of radius 10 having $n = 231$ locations. The diamond and rectangular grids, with a one-unit distance between adjacent locations, are used in § 5.3.

where $\kappa > 0$ is the shape parameter, $\phi > 0$ is the range parameter and K_κ is the modified Bessel function of the second kind of order κ. Our location pattern is constructed in the following way. The first 200 locations are drawn uniformly and randomly from an origin-centred square S_1 of side length $200^{1/2}$ units. For the next 200 locations, we scale the side length of the square S_1 by the factor $2^{1/2}$ to obtain a larger square S_2, and then draw the points uniformly and randomly from $S_2 \setminus S_1$. Subsequently, we always scale the side length of the previous square S_j by $2^{1/2}$ to obtain S_{j+1}, and then draw the same number of locations we already have on $S_{j+1} \setminus S_j$, thus doubling the number of points each time. This process is continued until we have obtained a total of 3200 locations. In the simulation we consider the sample sizes $n = 100 \times 2^j$ for $j = 1, \ldots, 5$, each time using the first n of the 3200 points, that is, all points inside the jth innermost square in Fig. 2(a). The six samples then correspond to nested samples of points and represent the increasing-domain asymptotic scheme implied by Assumption 5.

We expect any successful unmixing estimator $\hat{\Gamma}$ to satisfy $\hat{\Gamma}\Omega \approx I_p$ up to sign changes and row permutations. The minimum distance index (Ilmonen et al., 2010) is defined as

$$\mathrm{MDI}(\hat{\Gamma}) = (p-1)^{-1/2} \inf\{\|C\hat{\Gamma}\Omega - I_p\| : C \in \mathcal{C}\},$$

where $\mathcal{C}$ is the set of all matrices with exactly one nonzero element in each row and column and $\|\cdot\|$ is the Frobenius norm. The minimum distance index measures how close $\hat{\Gamma}\Omega$ is to the identity matrix up to scaling and the order and signs of its rows, and $0 \leqslant \mathrm{MDI}(\hat{\Gamma}) \leqslant 1$ with lower values indicating more efficient estimation. Moreover, for any $\hat{\Gamma}$ such that $n^{1/2}\,\mathrm{vect}(\hat{\Gamma} - I_p) \to N(0, \Sigma)$ for some limiting covariance matrix Σ, the transformed index $n(p-1)\mathrm{MDI}(\hat{\Gamma})^2$ converges to a limiting distribution $\sum_{i=1}^k \delta_i \chi_i^2$ where $\chi_1^2, \ldots, \chi_k^2$ are independent chi-squared random variables with one degree of freedom and $\delta_1, \ldots, \delta_k$ are the k nonzero eigenvalues of the matrix

$$\left(I_{p^2} - D_{p,p}\right) \Sigma \left(I_{p^2} - D_{p,p}\right)$$

with $D_{p,p} = \sum_{j=1}^p E^{jj} \otimes E^{jj}$; here E^{jj} is the $p \times p$ matrix with 1 as its (j,j)th element and the rest of its elements all equal to zero, and $\otimes$ is the usual tensor matrix product. In particular, the expected value of the limiting distribution is the sum of the limiting variances of the off-diagonal elements of $\hat{\Gamma}$. This provides us with a useful single-number summary to measure the asymptotic efficiency of the method, i.e., the mean value of $n(p-1)\mathrm{MDI}(\hat{\Gamma})^2$ over several replications.

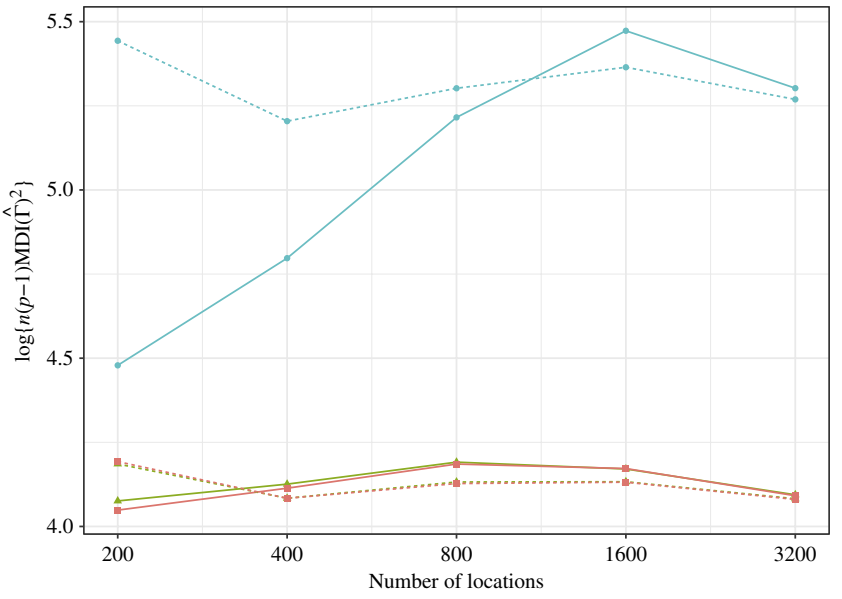

Fig. 3. The solid lines represent the mean values of $n(p-1)\text{MDI}(\hat{\Gamma})^2$ in the first simulation and the dashed lines correspond to the asymptotic approximations of the same quantities. The three local covariance matrices used are $B(1)$ (blue), $R(1,2)$ (green) and $\{B(1), R(1,2)\}$ (orange).

An argument given in the Supplementary Material can be used to show that our spatial blind source separation estimators are affine equivariant. More precisely, let $\hat{\Gamma}(I_p)$ be computed from $\{Z(s_i)\}_{i=1,\ldots,n}$ according to (3) and recall that $\hat{\Gamma}$ is computed from $\{X(s_i)\}_{i=1,\ldots,n}$ according to (3). Then we have $\hat{\Gamma} = \hat{\Gamma}(I_p)\Omega^{-1}$ up to sign changes and row permutations. In this sense, $\hat{\Gamma}\Omega$ is invariant with respect to the value of Ω. As the minimum distance index depends on $\hat{\Gamma}$ only through $\hat{\Gamma}\Omega$, throughout § 5 we may consider without loss of generality only the trivial mixing case of $\Omega = I_3$. Taking different Ω into consideration would give exactly the same results as described below.

Recall that the ball and ring kernels are defined, respectively, as $B(h)(s) = I(\|s\| \leqslant h)$ for fixed $h \geqslant 0$ and $R(h_1, h_2)(s) = I(h_1 \leqslant \|s\| \leqslant h_2)$ for fixed $h_2 \geqslant h_1 \geqslant 0$. We simulate 2000 replications for each sample size n and estimate the unmixing matrix in each case with three different choices of the local covariance matrix kernels: $B(1)$, $R(1,2)$ and $\{B(1), R(1,2)\}$, where the argument s is dropped and the brackets { } denote the joint diagonalization of the kernels enclosed. The latent covariance functions in Fig. 1(a) show that the dependencies of the last two fields die off rather quickly, and we would expect that very local information is already sufficient to separate the fields. Moreover, out of all one-unit intervals, the magnitudes of the three covariance functions differ the most from each other in the interval from 1 to 2, and we may reasonably assume that either $R(1,2)$ or $\{B(1), R(1,2)\}$ will be the most efficient choice.

The mean values of $n(p-1)\text{MDI}(\hat{\Gamma})^2$ over the 2000 replications are shown as solid lines in Fig. 3, with the dashed lines representing the asymptotic approximated values of the means, towards which they are expected to converge; see Propositions 4 and 7. As evidenced in Fig. 3, this is indeed what happens. For the reasons detailed in the previous paragraph, the kernel $R(1,2)$ is a considerably more efficient choice than $B(1)$. However, the ball kernel still carries some additional information over the ring kernel, as their joint diagonalization, $\{B(1), R(1,2)\}$, gives the best results out of the three choices, albeit marginally. As the main purpose of the current

simulation is to verify the limiting theorems and compare the different choices of kernels, the estimation accuracy of the sources is considered jointly, through the minimum distance index. However, as it is possible that some of the individual sources are more difficult to estimate than others, we have included a simulation study exploring individual component recovery in the Supplementary Material.

The previous investigation and Fig. 3 used only the expected value of the asymptotic distribution. In the Supplementary Material, we also plot the estimated densities of $n(p-1)\text{MDI}(\hat{\Gamma})^2$ for all local covariance matrices and a few selected sample sizes, and compare them with the density of the asymptotic approximation estimated from a sample of 100 000 random variables drawn from the corresponding distributions. Overall, the two densities match each other rather well, especially for the local covariance matrices involving the ring kernel. This shows that the asymptotic approximation to the distribution of $n(p-1)\text{MDI}(\hat{\Gamma})^2$ is good already for small sample sizes.

5.3. *The effect of range on the efficiency*

The second simulation explores the effect of the range of the latent fields on the asymptotically optimal choice of local covariance matrices. The comparisons between the estimators are made on the basis of the expected values of the asymptotic approximations to the distribution of $n(p-1)\text{MDI}(\hat{\Gamma})^2$, that is, using the equivalent of the dashed lines in Fig. 3, meaning that no randomness is involved in this simulation.

We consider three-variate random fields $X(s) = \Omega Z(s)$, where $\Omega = I_3$ and the latent fields have Matérn covariance functions with shape parameters $\kappa = 2, 1, 0.25$ and range parameter $\phi \in \{1.0, 1.1, 1.2, \ldots, 30.0\}$. The three covariance functions are shown for $\phi = 1$ in Fig. 1(b). The random field is observed at three different point patterns: diamond-shaped, rectangular and random, which was simulated once and held fixed throughout the study. The diamond-shaped point pattern has a radius of $m = 30$ and a total of $n = 1861$ locations, whereas the rectangular point pattern has a radius of $m = 15$ with a total of $n = 1891$ locations. In both patterns, the horizontal and vertical distance between adjacent locations is one unit, and examples of the two pattern types are shown in Figs. 2(b) and (c) with a radius of $m = 10$. A rectangular pattern with radius m is defined to have width $2m+1$ and height $m+1$. The random point pattern is generated by simply simulating $n = 1861$ points uniformly in the rectangle $(-30, 30) \times (-30, 30)$. We consider a total of eight different local covariance matrices: $B(r)$ and $R(r-1, r)$ for $r = 1, 3, 5$, as well as the joint diagonalizations of these sets, $\{B(1), B(3), B(5)\}$ and $\{R(0, 1), R(2, 3), R(4, 5)\}$.

The results of the simulation are displayed in Fig. 4, where the two joint diagonalizations are represented by the parameter r having value J. Recall that the lower the value on the vertical axis, the better that particular method is at estimating the three latent fields. The relative ordering of the different curves is very similar across all three plots, and it seems that the choice of location pattern does not have a large effect on the results. For all three patterns, the local covariance matrices with either $r = 1$ or $r = 3$ are the best choices for small values of ϕ, but their performance quickly deteriorates as ϕ increases. The opposite happens for the local covariance matrices with $r = 5$; they are among the worst-performing for small ϕ, but exhibit relative improvement with increasing ϕ. The joint diagonalization-based choices fall somewhere in-between and are never the best or the worst choice. However, they yield performance very close to the best choice in the right end of the ϕ range and are close to optimal in the left end. Thus, their use could be justified in practice as the safe choice. Comparing the two types of local covariance matrices, balls and rings, we observe that in the majority of cases the rings prove superior to the balls.

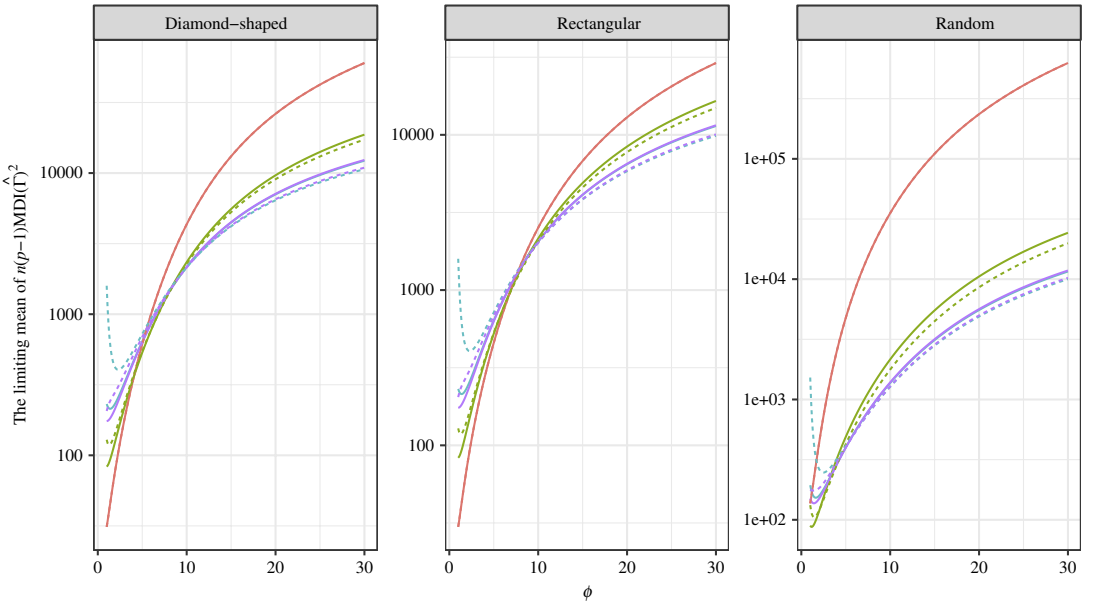

Fig. 4. Asymptotic approximate mean values of $n(p-1)\mathrm{MDI}(\hat{\Gamma})^2$ plotted as a function of the range of the latent Matérn random fields for the different choices of local covariance matrices in the second simulation. The solid and the dashed lines correspond to, respectively, the ball and ring kernels, and the value of the parameter r is indicated by the colour of the line: 1 (red), 3 (green), 5 (blue) and J (purple). The y-axis has a logarithmic scale.

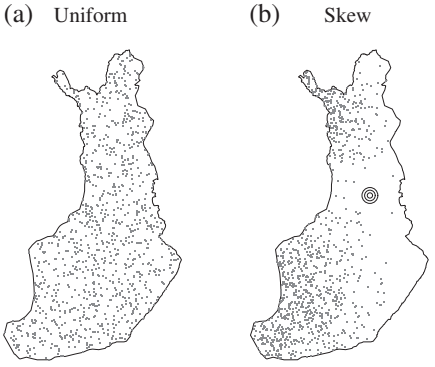

Fig. 5. The two fixed location patterns on a map of Finland: (a) uniform pattern; (b) skew pattern.

5.4. *Efficiency comparison*

To compare a larger number of local covariance matrices and their combinations, we simulate three-variate random fields $X(s) = \Omega Z(s)$, where $\Omega = I_3$ and the latent fields have Matérn covariance functions with shape parameters $\kappa = 6, 1, 0.25$ and range parameter $\phi = 20$, in kilometres. We consider two different fixed-location patterns fitted inside the map of Finland; see Fig. 5. The first pattern has the locations drawn uniformly from the map and the second pattern is drawn from a west-skew distribution. Both patterns have a total of $n = 1000$ locations, and to better distinguish the scale we have added three concentric circles with radii of 10, 20, and 30 kilometres in the empty area of the skew map.

We simulate a total of 2000 replications of the above scheme with the fixed maps. In each case we compute the minimum distance index values of the estimates obtained with the local covariance

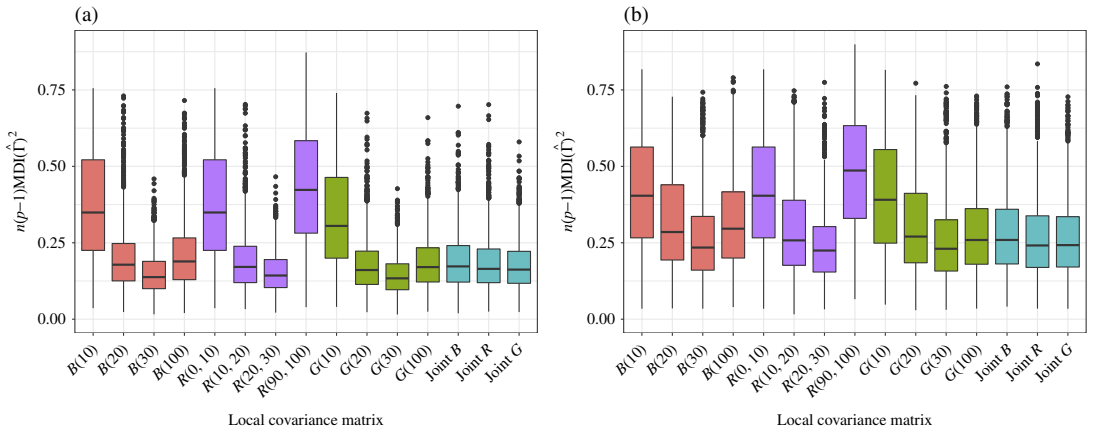

Fig. 6. Results of the efficiency study for (a) the uniform sampling design and (b) the skew design.

matrix kernels $B(r)$, $R(r-10, r)$ and $G(r)$, where $r = 10, 20, 30$ and 100, and the joint diagonalization of each of the three quadruplets $\{B(10), B(20), B(30), B(100\}$, $\{R(10), R(20), R(30), R(100\}$ and $\{G(10), G(20), G(30), G(100\}$, making a total of 15 estimators. The Gaussian kernel is parameterized as $G(r) \equiv \exp[-0.5\{\Phi^{-1}(0.95)s/r\}^2]$, where s is the distance and $\Phi^{-1}(x)$ is the quantile function of the standard normal distribution, so that $G(r)$ has 90% of its total mass in the ball of radius r around its centre. Thus, $G(r)$ can be considered a smooth approximation of $B(r)$. The larger-radius kernels $B(100)$, $R(90, 100)$ and $G(100)$ are included in the simulation to investigate what happens when we overestimate the dependency radius. The mean minimum distance index values for the 15 estimators are plotted in Fig. 6, and show that for both maps and all local covariance types, increasing the radius yields more accurate separation results all the way up to $r = 30$, but for $r = 100$ the results again worsen. This observation indicates that when a single local covariance matrix is used, the choice of the type and the radius are especially important, most likely requiring some expert knowledge about the study. However, this problem can be completely avoided if we use the joint diagonalization of several matrices. For both maps and all local covariance types, the joint diagonalization produces results comparable to the best individual matrices, even though the joint diagonalizations also include the bad choices of $r = 10, 20, 100$. We observe similar behaviour to that in the first and second simulation studies, where, in the absence of knowledge about the optimal choice, the joint diagonalization either is the most efficient choice or yields performance very close to that of the most efficient choice. Therefore, we recommend use of the joint diagonalization of scatter matrices with a sufficiently large variation of radii for the kernels.

Finally, a comparison of the two maps reveals that the relative behaviour of the estimators is roughly the same in both maps, but the estimation is generally more difficult in the skew map, reflected by the on-average higher minimum distance index values. This can be explained by the large number of isolated points that contribute no information to the estimation of the local covariance matrices, making the sample size essentially smaller than $n = 1000$.

6. Data application

To illustrate the benefit of jointly diagonalizing more than two scatter matrices from a practical point of view, we reconsider the moss data from the Kola project, available in the `R` package `StatDa` (Filzmoser, 2015) and described in Reimann et al. (2008), for example. The data consist

Table 1. *Maximal absolute correlations of different estimators with respect to the gold standard; all estimators used the empirical covariance matrix, and the distances for the scatters are given in kilometres*

Est	Scatters	IC1	IC2	IC3	IC4	IC5	IC6
1	$B(25)$	0.96	0.93	0.91	0.68	0.64	0.77
2	$B(75)$	0.98	0.98	0.92	0.96	0.91	0.63
3	$B(100)$	0.76	0.80	0.77	0.96	0.60	0.53
4	$R(0,25)$, $R(25,50)$, $R(50,75)$, $R(75,100)$	0.97	0.98	0.92	0.97	0.83	0.80
5	$R(0,10)$, $R(10,20)$, $R(20,30)$, $R(30,40)$, $R(40,50)$, $R(50,60)$, $R(60,70)$, $R(70,80)$	0.96	0.97	0.91	0.97	0.78	0.77

Est, estimator; IC, independent component.

of 594 samples of terrestrial moss collected at different sites in northern Europe on the borders of Norway, Finland and Russia. The corresponding map with sampling locations is shown in the Supplementary Material. The amounts of 31 chemical elements found in the moss samples have already been used as a spatial blind source separation example in Nordhausen et al. (2015), where the covariance matrix and $B(50)$ were simultaneously diagonalized. The goal of that analysis was to reveal interpretable components exhibiting clear spatial patterns. In Nordhausen et al. (2015), the radius of 50 km was carefully chosen by an expert and considered to be best among several choices of radius. The analysis found six meaningful components which could be used to distinguish underlying natural geological patterns from environmental pollution patterns. These six components had the six largest eigenvalues and are visualized in the Supplementary Material.

We show that the gold-standard components can be stably estimated without subject knowledge of the optimal radius by simply jointly diagonalizing a large enough collection of local covariance matrices. To address the compositional nature of the data, we follow the same data preparation steps as in Nordhausen et al. (2015) and then compute five competing spatial blind source separation estimates. The scatters we used in addition to the covariance matrix are detailed in Table 1. Using this approach, we identify the six components having the highest correlations, in absolute value, with the six main components identified in Nordhausen et al. (2015). Table 1 also gives the correlations of the six components.

The table shows that when using only two scatters, as in estimators 1, 2 and 3, some components cannot be easily found. However, if one jointly diagonalizes more than two scatters, the results are more stable and less dependent on the chosen distances of the scatters, as can be seen with estimators 4 and 5. This is illustrated using the gold standard and estimators 3 and 4 in Fig. A1 in the Appendix for the first two components. For completeness, the Supplementary Material presents all six components for the three estimators. The first two components represent, according to Nordhausen et al. (2015), areas with different types of industrial contamination, and Fig. A1 shows that the gold standard and estimator 4 agree quite well on these, but estimator 3 yields a different map. More precisely, the first component obtained from the gold standard and estimator 4 highlights a cluster of negative scores around the Monchegorsk and Apatity region, which reflects the mining and processing of alkaline deposits. This cluster is not revealed by estimator 3. Similarly, the second components are similar between the gold standard and estimator 4, but the component from estimator 3 differs from these two, especially for the sampling locations in Finland. Thus, using several scatters gives a more stable impression, whereas the maps can vary considerably when only two scatters are used, in which case subject-matter expertise becomes more important.

7. Discussion

The proposed method can be extended in multiple directions. The assumptions of Gaussian or stationary fields could be relaxed. The spatial and temporal blind source separation methods could be combined to obtain spatiotemporal blind source separation. If used for dimension reduction, estimators for the number of latent non-noise fields could be devised using strategies similar to those in Virta & Nordhausen (2020). Additionally, the combination of spatial blind source separation with univariate kriging and univariate modelling warrants investigation.

How to choose the local covariance matrices optimally is also of interest. This is still an open problem for temporal blind source separation methods, such as second-order blind identification (Belouchrani et al., 1997). Several strategies have been suggested (see, e.g., Tang et al., 2005), and many of them could be useful also in selecting the kernels in spatial blind source separation. The estimation accuracy of our proposed method is based on how well separated the eigenvalues of the matrices $M(f_0)^{-1/2}M(f_l)M(f_0)^{-1/2}$ $(l = 1, \ldots, k)$ are. Since the connection between the eigenvalues and the unknown covariance functions is complicated, our suggestion, backed up also by the simulations, is to stay on the safe side and use a large number of ring kernels jointly. However, including large numbers of unnecessary kernels can have the drawback of inducing some noise in the estimates. One way to remove the unneeded kernels would be to first obtain preliminary estimates for the latent fields using a large number of kernels jointly; then our asymptotic results could be used to select from a large collection of sets of kernels, the one which achieves the smallest value of $\delta_1 + \cdots + \delta_k$; see § 5.2. The final estimates could then be computed with this asymptotically optimal choice of kernels. A similar technique was used by Taskinen et al. (2016) in the context of temporal blind source separation.

Acknowledgement

The work of Nordhausen, Ruiz-Gazen and Virta was partly supported by the CRoNoS Cost action. Nordhausen was also supported by the Austrian Science Fund. Ruiz-Gazen acknowledges funding from the French National Research Agency under the Investissements d'Avenir programme. Joni Virta was supported by the Academy of Finland. The authors are very grateful for the comments of the referees, which helped to improve the manuscript considerably. Virta is also affiliated with Aalto University, Finland.

Supplementary material

Supplementary material available at *Biometrika* online includes the proofs of all theoretical results, an auxiliary simulation, and additional figures for the simulations and the data application.

Appendix

Notation

Let y and z be the $np \times 1$ vectors defined by $y_{(i-1)p+j} = Y_j(s_i)$ and $z_{(i-1)p+j} = Z_j(s_i)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Let $R = \mathrm{cov}(y)$ and $R_z = \mathrm{cov}(z)$. Let $e_b(p)$ be the bth base column vector of $\mathbb{R}^p$ for $b = 1, \ldots, p$. For $f : \mathbb{R}^d \to \mathbb{R}$ and $b, l = 1, \ldots, p$, let $T_{b,l}(f)$ be the $np \times np$ block matrix composed of n^2 blocks of size p^2, with block (i,j) equal to $f(s_i - s_j)(1/2)\{e_b(p)e_l(p)^{\mathrm{T}} + e_l(p)e_b(p)^{\mathrm{T}}\}$.

For $b \in \mathbb{N}$ let $\mathcal{D}(b) = \{1 + (i-1)(b+1) : i = 1, \ldots, b\}$. We remark that $\{\mathrm{vect}(M)_i : i \in \mathcal{D}(b)\} = \{M_{i,i} : i = 1, \ldots, b\}$ for a $b \times b$ matrix M. Let $\bar{\mathcal{D}}_b = \{1, \ldots, b^2\} \backslash \mathcal{D}_b$. We remark that $\{\mathrm{vect}(M)_i : i \in \bar{\mathcal{D}}(b)\} = \{M_{i,j} : i, j = 1, \ldots, b,\ i \neq j\}$ for a $b \times b$ matrix M. For $a \in \{1, \ldots, b^2\}$, let $I_b(a)$ and $J_b(a)$ be the

unique i and j from $\{1,\dots,b\}$ such that $a = b(i-1)+j$. For $i \in \{1,\dots,b\}$, let $d_b(i) = 1+(i-1)(b+1)$ and note that $\{\mathrm{vect}(M)_{d_b(i)} : i = 1,\dots,b\} = \{M_{i,i} : i = 1,\dots,b\}$ for a $b \times b$ matrix M. For a matrix M of size $b \times b$, recall that $\mathrm{diag}(M) = (M_{1,1},\dots,M_{b,b})^{\mathrm{T}}$ and that $\mathrm{tr}(M)$ denotes the trace of M.

Expression for the matrix $V(f,f_0)$ in Proposition 3

Let $f, g : \mathbb{R}^d \to \mathbb{R}$. Using the notation above, let $\Sigma(f)$ and $\Sigma(f,g)$ be the $p^2 \times p^2$ matrices defined, for $i = (s-1)p+t$ and $j = (u-1)p+v$ with $s,t,u,v \in \{1,\dots,p\}$, by

$$\Sigma(f)_{i,j} = 2n^{-1}\mathrm{tr}\{RT(f)_{s,t}RT(f)_{u,v}\}, \qquad \Sigma(f,g)_{i,j} = 2n^{-1}\mathrm{tr}\{RT(f)_{s,t}RT(g)_{u,v}\}. \tag{A1}$$

Let

$$V(f,g) = \begin{Bmatrix} \Sigma(f) & \Sigma(f,g) \\ \Sigma(g,f) & \Sigma(g) \end{Bmatrix}.$$

Then $V(f,f_0)$ is equal to $V(f,g)$ for $g = f_0$.

Expression for the matrix F_1 in Proposition 4

From Assumption 8, there exists $n_0 \in \mathbb{N}$ such that for $n \geqslant n_0$ the diagonal elements of $\Omega^{-1}M(f)\Omega^{-\mathrm{T}}$ are strictly decreasing. Write these diagonal elements as $\lambda_1 > \cdots > \lambda_p$. Using the notation defined at the beginning of this Appendix, for $n \geqslant n_0$ let A, B, C and D be, respectively, the $p^2 \times p^2$, $p^2 \times p^2$, $p \times p^2$ and $p \times p^2$ matrices defined by

$$A_{i,j} = \begin{cases} -1/2, & i = j \in \mathcal{D}(p), \\ -\lambda_{I_p(i)}\{\lambda_{I_p(i)} - \lambda_{J_p(i)}\}^{-1}, & i = j \notin \mathcal{D}(p), \\ 0, & \text{otherwise}, \end{cases} \quad B_{i,j} = \begin{cases} \{\lambda_{I_p(i)} - \lambda_{J_p(i)}\}^{-1}, & i = j \notin \mathcal{D}(p), \\ 0, & \text{otherwise}, \end{cases}$$

$$C_{i,j} = \begin{cases} -\lambda_i, & j = d_p(i), \\ 0, & \text{otherwise}, \end{cases} \qquad D_{i,j} = \begin{cases} 1, & j = d_p(i), \\ 0, & \text{otherwise}. \end{cases}$$

Let

$$G = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

Let $M_{\Omega^{-1}}$ and $\bar{M}_{\Omega^{-1}}$ be, respectively, the $p^2 \times p^2$ and $(p^2+p) \times (p^2+p)$ matrices defined by

$$(M_{\Omega^{-1}})_{a,b} = \begin{cases} (\Omega^{-1})_{J_p(b),J_p(a)}, & I_p(a) = I_p(b), \\ 0, & I_p(a) \neq I_p(b), \end{cases} \qquad \bar{M}_{\Omega^{-1}} = \begin{pmatrix} M_{\Omega^{-1}} & 0 \\ 0 & I_p \end{pmatrix}. \tag{A2}$$

Let $\tilde{V}(f)$ be defined as $V(f_0,f)$, but with R replaced by R_z. Then for $n \geqslant n_0$, F_1 is defined as

$$F_1 = \bar{M}_{\Omega^{-1}} G \tilde{V}(f) G^{\mathrm{T}} \bar{M}_{\Omega^{-1}}^{\mathrm{T}}.$$

Expression for the matrix F_k in Proposition 7

Let $D(f) = \Omega^{-1}M(f)\Omega^{-\mathrm{T}}$. For a diagonal matrix Λ, let $\Lambda_r = \Lambda_{r,r}$. With the notation of Assumption 9, let $A_0, A_1, \dots, A_k$ and B be $p^2 \times p^2$ matrices defined, for $n \geqslant n_0$, by

$$A_{0,i,j} = \begin{cases} -1/2, & i = j \in \mathcal{D}(p), \\ -\sum_{l=1}^{k}\{D(f_l)_{I_p(i)} - D(f_l)_{J_p(i)}\}D(f_l)_{I_p(i)}, & i = j \notin \mathcal{D}(p), \\ 0 & \text{otherwise}, \end{cases}$$

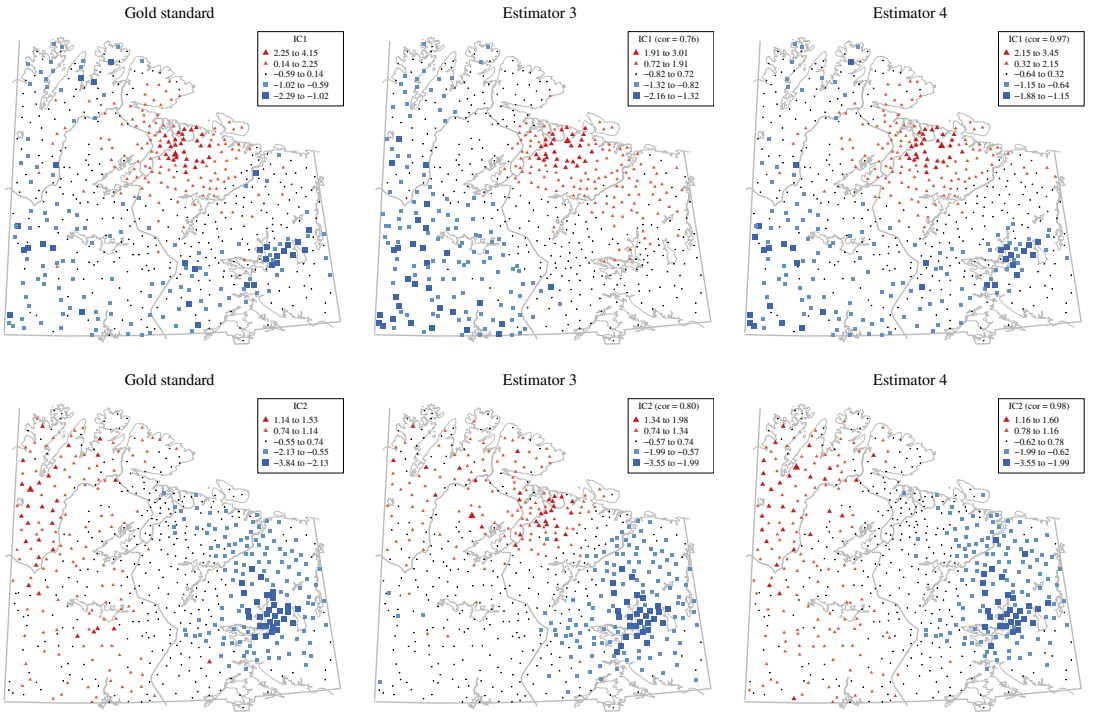

Fig. A1. The first two independent components from the gold standard and estimators 3 and 4.

$$A_{l,i,j} = \begin{cases} D(f_l)_{I_p(i)} - D(f_l)_{J_p(i)}, & i = j \notin \mathcal{D}(p), \\ 0, & \text{otherwise} \end{cases}$$

for $l = 1, \ldots, k$, and

$$B_{i,j} = \begin{cases} 1, & i = j \in \mathcal{D}(p), \\ \left[\sum_{l=1}^{k} \{D(f_l)_{I_p(i)} - D(f_l)_{J_p(i)}\}^2 \right]^{-1}, & i = j \notin \mathcal{D}(p), \\ 0, & \text{otherwise.} \end{cases}$$

Let G be the $p^2 \times (k+1)p^2$ matrix defined by $G = B(A_0, A_1, \ldots A_k)$ for $n \geqslant n_0$. Let $M_{\Omega^{-1}}$ be as in (A2). Let $\tilde{V}(f_1, \ldots, f_k)$ be the $(k+1)p^2 \times (k+1)p^2$ matrix composed of $(k+1)^2$ blocks of size $p^2 \times p^2$ with block $((i+1),(j+1))$ defined similarly to $\Sigma(f_i, f_j)$ in (A1), but with R replaced by R_z. Then for $n \geqslant n_0$, F_k is defined as

$$F_k = M_{\Omega^{-1}} G \tilde{V}(f_1, \ldots, f_k) G^{\mathrm{T}} M_{\Omega^{-1}}^{\mathrm{T}}.$$

Map for data application

The map for the data example in § 6 is shown in Fig. A1.

References

Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F. & Moulines, E. (1997). A blind source separation technique using second-order statistics. *IEEE Trans. Sig. Proces.* **45**, 434–44.

Clarkson, D. B. (1988). Remark AS R71: A remark on algorithm AS 211. The F-G diagonalization algorithm. *Appl. Statist.* **37**, 147–51.

COMON, P. & JUTTEN, C. (2010). *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Oxford: Academic Press.
CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. New York: John Wiley & Sons, 2nd ed.
DE IACO, S., MYERS, D. E., PALMA, M. & POSA, D. (2013). Using simultaneous diagonalization to identify a space–time linear coregionalization model. *Math. Geosci.* **45**, 69–86.
DUDLEY, R. M. (2002). *Real Analysis and Probability*. Cambridge: Cambridge University Press.
EDDELBUETTEL, D. & SANDERSON, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Comp. Statist. Data Anal.* **71**, 1054–63.
EMERY, X. (2010). Iterative algorithms for fitting a linear model of coregionalization. *Comp. Geosci.* **36**, 1150–60.
FILZMOSER, P. (2015). *StatDA: Statistical Analysis for Environmental Data*. R package version 1.6.9.
GELFAND, A. E., SCHMIDT, A. M., BANERJEE, S. & SIRMANS, C. F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *Test* **13**, 263–312.
GENTON, M. G. & KLEIBER, W. (2015). Cross-covariance functions for multivariate geostatistics. *Statist. Sci.* **30**, 147–63.
GOULARD, M. & VOLTZ, M. (1992). Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix. *Math. Geol.* **24**, 269–86.
ILLNER, K., MIETTINEN, J., FUCHS, C., TASKINEN, S., NORDHAUSEN, K., OJA, H. & THEIS, F. J. (2015). Model selection using limiting distributions of second-order blind source separation algorithms. *Sig. Proces.* **113**, 95–103.
ILMONEN, P., NORDHAUSEN, K., OJA, H. & OLLILA, E. (2010). A new performance index for ICA: Properties, computation and asymptotic analysis. In *Latent Variable Analysis and Signal Separation*. Cham: Springer, pp. 229–36.
MATILAINEN, M., NORDHAUSEN, K. & OJA, H. (2015). New independent component analysis tools for time series. *Statist. Prob. Lett.* **105**, 80–7.
MIETTINEN, J., ILLNER, K., NORDHAUSEN, K., OJA, H., TASKINEN, S. & THEIS, F. (2016). Separation of uncorrelated stationary time series using autocovariance matrices. *J. Time Ser. Anal.* **37**, 337–54.
MIETTINEN, J., NORDHAUSEN, K., OJA, H. & TASKINEN, S. (2012). Statistical properties of a blind source separation estimator for stationary time series. *Statist. Prob. Lett.* **82**, 1865–73.
MIETTINEN, J., NORDHAUSEN, K., OJA, H. & TASKINEN, S. (2014). Deflation-based separation of uncorrelated stationary time series. *J. Mult. Anal.* **123**, 214–27.
MIETTINEN, J., NORDHAUSEN, K. & TASKINEN, S. (2017). Blind source separation based on joint diagonalization in R: The packages JADE and BSSasymp. *J. Statist. Software* **76**, 1–31.
NORDHAUSEN, K. (2014). On robustifying some second order blind source separation methods for nonstationary time series. *Statist. Papers* **55**, 141–56.
NORDHAUSEN, K. & OJA, H. (2018). Independent component analysis: A statistical perspective. *WIREs Comp. Statist.* **10**, e1440.
NORDHAUSEN, K., OJA, H., FILZMOSER, P. & REIMANN, C. (2015). Blind source separation for spatially correlated compositional data. *Math. Geosci.* **47**, 753–70.
R DEVELOPMENT CORE TEAM (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.
REIMANN, C., FILZMOSER, P., GARRETT, R. & DUTTER, R. (2008). *Statistical Data Analysis Explained. Applied Environmental Statistics with R*. Chichester: Wiley.
RIBEIRO JR, P. J. & DIGGLE, P. J. (2016). *geoR: Analysis of Geostatistical Data*. R package version 1.7-5.2.
TANG, A. C., LIU, J.-Y. & SUTHERLAND, M. T. (2005). Recovery of correlated neuronal sources from EEG: The good and bad ways of using SOBI. *Neuroimage* **28**, 507–19.
TASKINEN, S., MIETTINEN, J. & NORDHAUSEN, K. (2016). A more efficient second order blind identification method for separation of uncorrelated stationary time series. *Statist. Prob. Lett.* **116**, 21–6.
VIRTA, J. & NORDHAUSEN, K. (2020). Determining the signal dimension in second order source separation. *Statist. Sinica* to appear, DOI: 10.5705/ss.202018.0347.
WACKERNAGEL, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*. Berlin: Springer, 3rd ed.
XIE, T. & MYERS, D. E. (1995). Fitting matrix-valued variogram models by simultaneous diagonalization (Part I: Theory). *Math. Geol.* **27**, 867–75.
XIE, T., MYERS, D. E. & LONG, A. E. (1995). Fitting matrix-valued variogram models by simultaneous diagonalization (Part II: Application). *Math. Geol.* **27**, 877–88.
ZHANG, H. (2007). Maximum-likelihood estimation for multivariate spatial linear coregionalization models. *Environmetrics* **18**, 125–39.

[*Received on* 19 *December* 2018. *Editorial decision on* 22 *August* 2019]