Spatial Blind Source Separation

François Bachoc, Marc G. Genton, Klaus Nordhausen, Anne Ruiz-Gazen, Joni Virta

Spatial Blind Source Separation

François BACHOC

Institut de Mathématiques de Toulouse,
Université Paul Sabatier, 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
CSTAT, University of Technology,
Wiedner Hauptstr. 7, A-1040 Vienna, Austria
klaus.nordhausen@tuwien.ac.at

Anne RUIZ-GAZEN
Toulouse School of Economics, University Toulouse Capitole,
21 allée de Brienne, 31000 Toulouse, France
anne.ruiz-gazen@tse-fr.eu

Joni VIRTA
Aalto University, Helsinki, Finland
joni.virta@aalto.fi

December 21, 2018

Abstract

Recently a blind source separation model was suggested for spatial data together 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 limiting properties and merits of the novel estimator are verified in simulation studies. A real data example illustrates the method.

Keywors: joint diagonalisation, limiting distribution, multivariate random fields, spatial scatter matrices.

1 Introduction

There is an abundance of multivariate data measured at spatial locations $s_1, \ldots, s_n$ in a domain $\mathcal{S}^d$. Considering the spatial dependence when modelling such data is essential because measurements taken closer to each other tend to be more similar than measurements taken further apart. Hence, the spatial cross-covariance matrix C is crucial for such modelling. Let $X(s) = \{X_1(s), \ldots, X_p(s)\}^{\mathrm{T}}$ be a p-variate random field over the d-dimensional domain $\mathcal{S}^d$, where $^{\mathrm{T}}$ denotes the transpose operator. The cross-covariance matrix functional is then

$$C(s_1, s_2) = \mathrm{cov}\{X(s_1), X(s_2)\} := \{C_{i,j}(s_1, s_2)\}_{i,j=1}^{p}.$$

In a recent extensive review, Genton and Kleiber (2015) discussed different approaches to define such functionals and gave a list of properties and conventions that they should verify. In this paper, we discuss isotropic fields which are both second-order stationary and invariant under rotations. This means that the direction of the difference $h = s_1 - s_2$ does not matter and only the distance between the locations determines the covariances, i.e.

$$C_{i,j}(s_1, s_2) = C_{i,j}(\|h\|),$$

where $\|\cdot\|$ denotes the Euclidean norm. Such fields are very popular in practice as these working assumptions make methods more tractable.

As Genton and 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, and in textbooks such as Wackernagel (2003) usually the following two popular models are described.

In the intrinsic correlation model it is assumed that the 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 positive 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 then has the form

$$C(h) = \sum_{g=1}^{r} \rho_g(h) T_g,$$

for some positive integer $r \leq p$ with all the ρ_g's being univariate spatial correlation functions and T_g's being positive definite $p \times p$ matrices often called coregionalization matrices. Hence with $r = 1$ this reduces to the intrinsic correlation model. The linear model of coregionalization implies a symmetric covariance matrix.

Estimation in the linear model of coregionalization is discussed in several papers. Goulard and Voltz (1992) focused on the coregionalization matrices using an

iterative algorithm where the spatial correlation functions are assumed to be known. The algorithm was extended in Emery (2010). Assuming Gaussian random fields, an expectation-maximisation algorithm was suggested in Zhang (2007) and a Bayesian approach was considered in Gelfand et al. (2004).

Blind source separation is a general multivariate approach to construct models based on latent components with desirable properties. It was established as an independent component analysis for independent and identically distributed data and for stationary and non-stationary time series. Comon and Jutten (2010) and Nordhausen and Oja (2018) are general recent references. The goal in blind source separation is then to estimate the latent components which often allows a dimension reduction and to continue the analysis by fitting univariate models.

Recently Nordhausen et al. (2015) suggested a spatial blind source separation model where the latent components could be assumed to be uncorrelated or independent, allowing, for example, for univariate kriging, which is otherwise quite demanding in higher dimensions (Furrer and Genton, 2011).

In this paper, in Section 2 we explain the spatial blind source separation model in detail and relate it to the coregionalization model. In Section 3 we derive the properties of the estimator suggested in Nordhausen et al. (2015) and in Section 4 we suggest a novel estimator and present its statistical properties. We demonstrate the theoretical properties and merits of the novel estimator in a simulation study in Section 5 and we conclude with a data example in Section 6. All proofs and technical details are collected in the Appendix.

2 Spatial blind source separation model

Multivariate spatial models are often defined starting with their covariance function. In the spatial blind source separation model suggested in Nordhausen et al. (2015) it is more natural to start with the components. They proposed that the observed p-variate random field is a linear combination of p univariate random fields which are all uncorrelated; for convenience often the stronger assumption is made that the p random fields are independent.

Denoting the p unobservable random fields as $Z(s) = \{Z_1(s), \ldots, Z_p(s)\}^{\mathrm{T}}$, the observable p-variate vector $X(s)$ at a location s is a linear mixture of the underlying latent vector $Z(s)$

$$X(s) = \Omega Z(s),$$

where Ω is a $p \times p$ full rank mixing matrix with columns $\omega_1, \ldots, \omega_p$. In the spatial blind source separation model, the following assumptions are made:

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

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

Assumption 3. $\mathrm{cov}\{Z(s_1), Z(s_2)\} = \mathrm{E}\{Z(s_1)Z(s_2)^{\mathrm{T}}\} = D(s_1, s_2)$*, where* D *is a diagonal matrix.*

Assumption 1 is made for convenience and can easily be replaced by assuming a constant unknown mean (see Lemma A.2 in the Appendix). Assumption 2 says that the components of $Z(s)$ are uncorrelated. Furthermore, it implies that the variances of the components of $Z(s)$ are one, which prevents identifiability issues and comes without loss of generality. Assumption 3 says that there is also no spatial cross-dependence between the components. However, the model is not well-defined. The order of the latent fields and also their signs can be changed. This is common for all blind source separation approaches and is not considered as a problem in practice.

The goal in spatial blind source separation is to recover the latent p random fields and then work on them univariately. For centered intrinsic stationary fields, Nordhausen et al. (2015) suggested the use of some local covariance matrix:

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

with $f_h(s_i - s_j) = I(\|s_i - s_j\| < h)$ for a positive constant h where $I(\cdot)$ denotes the indicator function. Choosing $f_0(s) = I(s = 0)$ yields the covariance matrix

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

In the following, we consider a more general class of functions $f : \mathbb{R}^d \to \mathbb{R}$ called kernels and consider $M(f) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} f(s_i - s_j) \mathrm{E}\{X(s_i) X(s_j)^{\mathrm{T}}\}$. We denote by $A_{k,l}$ (resp. u_k) the (k, l) (resp. k) element of a matrix A (resp. vector u).

The problem in spatial blind source separation can be formulated as the search for a $p \times p$ unmixing matrix Γ such that $\Gamma X(s) = Z(s)$, where the equality is up to signs and order of the components. Following Nordhausen et al. (2015), an unmixing functional can be defined as

Definition 1. *For any function $f : \mathbb{R}^d \to \mathbb{R}$, the unmixing matrix functional $\Gamma(f)$ is defined as the functional which jointly diagonalizes $M(f)$ and $M(f_0)$ in the following way*

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

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

The unmixing matrix $\Gamma(f)$ can then be found using the generalized eigenvalue-eigenvector theory and is well defined if the eigenvalues of $M(f_0)^{-1} M(f)$ are distinct.

Remark 1. *The covariance function $C_X(h)$ resulting from a spatial blind source separation model is always symmetric. It can be written also as*

$$C_X(h) = \sum_{g=1}^{p} \rho_{Z,g}(h) T_g,$$

where $\rho_{Z,g}(h)$ is the correlation function of the gth component of Z and $T_g = C_{Z,g}(0)^{-1} \omega_g \omega_g^{\mathrm{T}} = \omega_g \omega_g^{\mathrm{T}}$ with $C_{Z,g}(0)$ denoting the covariance function of the gth component of Z. Thus the model is a special case of the linear model of coregionalization with $r = p$ and where all coregionalization matrices T_g, $g = 1, \ldots, p$, are rank one matrices.

3 Estimation and asymptotic properties

Let $s_1, \ldots, s_n$ be n observation points in $\mathcal{S}^d$. Consider the processes $X(s)$ and $Z(s)$ as defined previously with $X(s) = \Omega Z(s)$ where Ω is a full rank mixing matrix and suppose that Assumptions 1 to 3 hold. Consider a function $f : \mathbb{R}^d \to \mathbb{R}$ and define a local covariance matrix estimator by

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

The covariance estimator is

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

Definition 2. *The unmixing matrix estimator $\widehat{\Gamma}(f)$ jointly diagonalizes $\widehat{M}(f_0)$ and $\widehat{M}(f)$ in the following way*

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

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

Let $\mathrm{cov}\{Z_k(s_i), Z_k(s_j)\} = K_k(s_i - s_j) = D(s_i, s_j)_{k,k}$, where K_k denotes the covariance function of Z_k, for $k = 1, \ldots, p$. Asymptotic results can be derived for the previous estimators assuming that Assumptions 1 to 3 hold together with the following assumptions:

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

Assumption 5. *There exists a fixed $\Delta > 0$ so that, for all $n \in \mathbb{N}$ and, for all $i \neq j$, $i, j = 1, \ldots, n$, $\|s_i - s_j\| \geq \Delta$;*

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

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

Assumption 7. *There exist fixed $A > 0$ and $\alpha > 0$ such that, for all $x \in \mathbb{R}^d$ and, for all $k = 1, \ldots, p$,*

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

Assumption 5 implies that $\mathcal{S}^d$ is unbounded as $n \to \infty$, which means that we address 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 $B(h)(s) = I(\|s\| \leq h)$ with fixed $h \geq 0$ and $R(h_1, h_2)(s) = I(h_1 \leq \|s\| \leq h_2)$ with fixed $h_2 \geq h_1 \geq 0$.

Proposition 1 gives (i) the consistency of the estimator $\widehat{M}(f)$ and (ii) the joint asymptotic distribution of $\{\widehat{M}(f_0), \widehat{M}(f)\}$, where f satisfies Assumption 7. Proposition 2 gives the joint asymptotic distribution of the estimators $\widehat{\Gamma}(f)$ and $\widehat{\Lambda}(f)$. The asymptotic normality results are obtained by considering 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). In both propositions, we prove that the distance between the sequence of the estimators' distribution and some Gaussian random variable decreases to zero when n increases. The asymptotic variances are detailed in the Appendix.

Proposition 1. *Suppose $n \to \infty$ and Assumptions 1 to 6 hold and let $f : \mathbb{R}^d \to \mathbb{R}$ satisfy Assumption 7.*

(i) Then $\widehat{M}(f) - M(f) \to 0$ in probability when $n \to \infty$.

(ii) Let $W(f)$ be the vector of size $p^2 \times 1$, defined for $i = (a-1)p + b$, $a, b \in \{1, \ldots, p\}$, by

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

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

$$d_w\{Q_n, \mathcal{N}(0, V(f, f_0))\} \to 0,$$

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

Taking $f(s) = I(\|s\| \leq h)$ with $h > 0$ in Proposition 1 gives the asymptotic properties of the method proposed in Nordhausen et al. (2015).

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 and for a matrix A of size $k \times k$, let $\mathrm{diag}(A) = (A_{1,1}, \ldots, A_{k,k})^{\mathrm{T}}$.

Remark 2. *The previous proposition remains valid when centering the process X by $\bar{X} = n^{-1}\sum_{i=1}^n X(s_i)$. Indeed, we prove in Lemma A.2 of the Appendix that the difference between the centered estimator and $\widehat{M}(f)$ is of order $O_p(n^{-1})$.*

In the next proposition, we let $L^{-\mathrm{T}} = (L^{-1})^{\mathrm{T}}$ for an invertible matrix L.

Proposition 2. *Assume the same assumptions as in Proposition 1. Assume also that there exists $\delta > 0$ so that for all $n \in \mathbb{N}$, for every pair $i \neq j$, $i, j = 1, \ldots, p$, $|[\Omega^{-1}M(f)\Omega^{-\mathrm{T}}]_{i,i} - [\Omega^{-1}M(f)\Omega^{-\mathrm{T}}]_{j,j}| \geq \delta$. Assume also that the diagonal elements of $\Omega^{-1}M(f)\Omega^{-\mathrm{T}}$ are in decreasing order. Let Q_n be the distribution of*

$$n^{1/2}\begin{pmatrix} \mathrm{vect}\left\{\widehat{\Gamma}(f) - \Gamma(f)\right\} \\ \mathrm{diag}\left\{\widehat{\Lambda}(f) - \Lambda(f)\right\} \end{pmatrix}.$$

Then, we can choose $\widehat{\Gamma}(f)$ and $\widehat{\Lambda}(f)$ in Definition 2 so that when $n \to \infty$

$$d_w\{Q_n, \mathcal{N}(0, F)\} \to 0,$$

where details concerning the matrix F are given in the Appendix.

The performance of this estimator depends on the choice of $M(f)$ that should be chosen so that $\widehat{\Lambda}(f)$ 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 (see, e.g., Belouchrani et al. (1997); Miettinen et al. (2014); Matilainen et al. (2015); Miettinen et al. (2016)) and this concept will be applied to the spatial context in the following section.

4 Improving the estimation of the spatial blind source separation model by jointly diagonalizing more than two matrices

Assume that we have scatter matrices $V_0, V_1, \ldots, V_k$ which are $p \times p$ symmetric positive definite affine equivariant matrices in the sense that $V_i\{AX(s)+b\} = AV_i\{X(s)\}A^{\mathrm{T}}$, for all $p \times p$ full rank matrices A, all p-vectors b and all $i = 0, \ldots, k$. Assume also that in the spatial blind source separation model under consideration, we have $V_i\{Z(s)\}$ diagonal for $i = 0, \ldots, k$. Then the spatial blind source separation unmixing matrix estimation can be formulated via the following steps:

1. Calculate the standardized field $X_{\mathrm{st}}(s) = V_0\{X(s)\}^{-1/2}X(s)$;
2. Find the $p \times p$ orthogonal matrix U such that $V_i\{UX_{\mathrm{st}}(s)\}$ is diagonal for $i = 1, \ldots, k$;
3. The unmixing matrix is then $\Gamma = UV_0\{X(s)\}^{-1/2}$.

When $k = 1$, we recover the particular case of two matrices from Section 3 and the joint diagonalization can be exact. However, if $k > 1$, then for finite data, the joint diagonalization may be performed only approximately and the problem is usually formulated as

$$\underset{UU^{\mathrm{T}}=I_p}{\operatorname{argmin}} \sum_{i=1}^{k} \|\mathrm{off}\,[UV_i\{X_{st}(s)\}U^{\mathrm{T}}]\,\|_F^2,$$

where $\mathrm{off}(A)$ is the matrix obtained by replacing the diagonal elements of A by zeros and $\|A\|_F$ is the Frobenius norm. From experience in time series blind source separation (see for example Miettinen et al., 2016), usually the symmetrization of several matrices gives a better separation than those based on two matrices only.

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.

Proposition 3. *Suppose Assumptions 1 to 6 hold. Let $k \in \mathbb{N}$ be fixed. Let $f_1, \ldots, f_k : \mathbb{R}^d \to \mathbb{R}$ satisfy Assumption 7. Assume that there exists a fixed $\delta > 0$ so that for all $n \in \mathbb{N}$, for every pair $i \neq j$, $i, j = 1, \ldots, p$, there exists $l = 1, \ldots, k$ such that $|[\Omega^{-1}M(f_l)\Omega^{-\mathrm{T}}]_{i,i} - [\Omega^{-1}M(f_l)\Omega^{-\mathrm{T}}]_{j,j}| \geq \delta$. Let $\widehat{\Gamma} = \widehat{\Gamma}\{\widehat{M}(f_0), \widehat{M}(f_1), \ldots, \widehat{M}(f_k)\}$*

be such that

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

Then we can choose $\widehat{\Gamma}$ so that $\widehat{\Gamma} \to \Omega^{-1}$ in probability when n goes to infinity.

Proposition 4. *Assume the same assumptions as in Proposition 3. Let $(\widehat{\Gamma}_n)_{n\in\mathbb{N}}$ be any sequence of $p\times p$ matrices so that for any $n \in \mathbb{N}$, $\widehat{\Gamma}_n = \widehat{\Gamma}_n\{\widehat{M}(f_0), \widehat{M}(f_1), \ldots, \widehat{M}(f_k)\}$ satisfies (1). Then, there exists a sequence of permutation matrices (P_n) and a sequence of diagonal matrices (D_n), with diagonal components in $\{-1, 1\}$ so that the distribution Q_n of $n^{1/2}(\check{\Gamma}_n - \Omega^{-1})$ with $\check{\Gamma}_n = D_nP_n\widehat{\Gamma}_n$, satisfies as $n \to \infty$*

$$d_w\{Q_n, \mathcal{N}(0, F)\} \to 0,$$

where details concerning the matrix F are given in the Appendix.

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

5 Simulations

In the following we use simulated data to verify our asymptotic results and to compare the efficiencies of the different local covariance estimates under a varying set of spatial models. All simulations are performed in R (R Core Team, 2015) with the help of the packages *geoR* (Ribeiro Jr and Diggle, 2016), *JADE* (Miettinen et al., 2017), *Rcpp* (Eddelbuettel and François, 2011) and *RcppArmadillo* (Eddelbuettel and Sanderson, 2014).

5.1 Convergence of the unmixing matrix estimator to the limit

We start with a simple simulation to establish the validity of the limiting distribution of the unmixing matrix estimator $\widehat{\Gamma}(f)$ for different kernels f and to obtain some preliminary comparative results between the proposed estimators. We consider a centered, 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, range) parameters $(\kappa, \phi) \in \{(6, 1.2), (1, 1.5), (0.25, 1)\}$, which correspond to the left panel in Figure 1. We recall that the Matérn covariance (correlation) function is defined by

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

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. For simplicity, we use as the underlying point

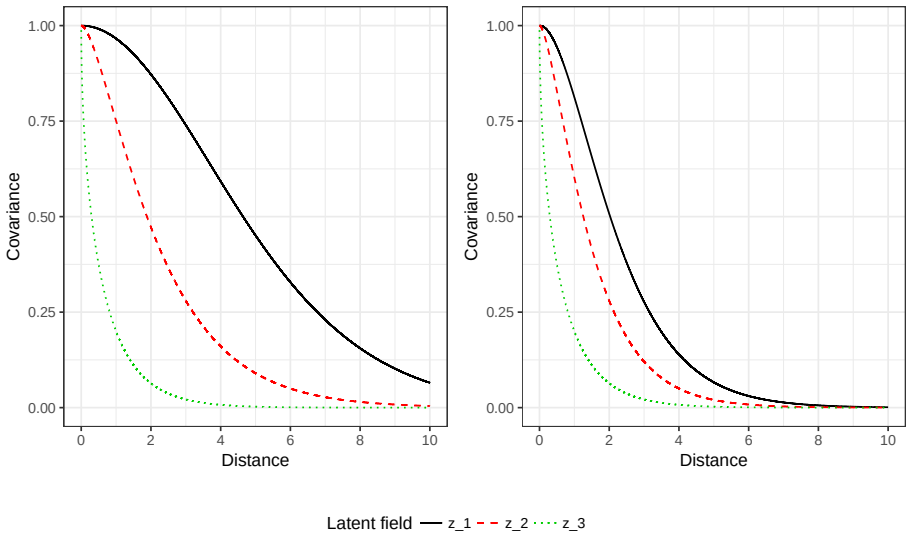

Figure 1: Matérn covariance functions of the three latent fields in the first and second simulation settings. The parameter vectors (κ, ϕ) of the three fields equal, $(6, 1.2), (1, 1.5), (0.25, 1)$ and $(2, 1), (1, 1), (0.25, 1)$, respectively.

pattern the diamond-shaped grid with a varying radii $m = 5, 10, \ldots, 50$, as depicted on the left-hand side of Figure 2 for $m = 10$. The number of locations as a function of the radius is $n = 2m(m + 1) + 1$.

As our spatial blind source separation estimators are affine equivariant, we consider without loss of generality the trivial mixing case $\Omega = I_3$. Any successful unmixing estimator $\widehat{\Gamma}$ is thus assumed to satisfy $\widehat{\Gamma} \approx I_p$ already for finite sample sizes (up to sign changes and row permutations). To account for the unidentifiability of the order and signs of the rows of $\widehat{\Gamma}$, we quantify our goal with the minimum distance index (Ilmonen et al., 2010b),

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

where $\mathcal{C}$ is the set of all matrices with exactly one non-zero element in each row and column. MDI measures how close its argument is to the identity matrix up to scaling, order and signs of its rows and $0 \leq \mathrm{MDI}(\widehat{\Gamma}) \leq 1$ with lower values indicating more efficient estimation. Moreover, for any $\widehat{\Gamma}$ such that $n^{1/2}\mathrm{vect}(\widehat{\Gamma} - I_p) \to \mathcal{N}(0, \Sigma)$ for some limiting covariance matrix Σ, the transformed index $n(p-1)\mathrm{MDI}(\widehat{\Gamma})^2$ 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 non-zero eigenvalues of the matrix,

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

where $D_{p,p} = \sum_{j=1}^{p} E^{jj} \otimes E^{jj}$ and E^{jj} is the $p \times p$ matrix with one as its (j, j)th

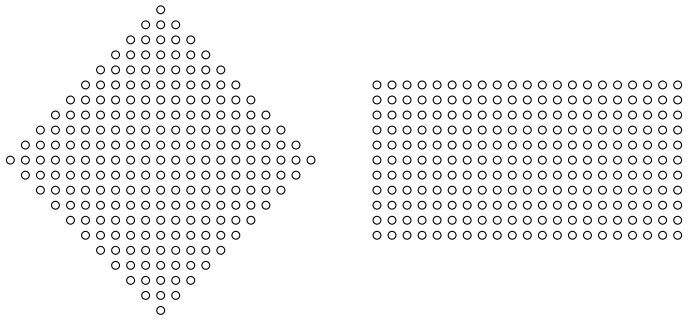

Figure 2: Diamond and rectangle grids of radius 10. The numbers of locations are respectively $n = 221, 231$ and the distance between two neighbouring locations is one unit.

element and the rest of the elements equal 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 $\widehat{\Gamma}$. This provides us with a useful single-number summary to measure the asymptotic efficiency of the method, the mean value of $n(p-1)\mathrm{MDI}(\widehat{\Gamma})^2$ over several replications.

Recall that the ball and ring kernels are defined as $B(h)(s) = I(\|s\| \leq h)$ and $R(h_1, h_2)(s) = I(h_1 \leq \|s\| \leq h_2)$ for fixed $h \geq 0$ and $h_2 \geq h_1 \geq 0$. We simulate 2000 replications of the field for each sample size n and estimate the unmixing matrix in each case with three different choices for 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 inside. The latent covariance functions in the upper half of Figure 1 show that the dependencies of the last two series die off rather quickly, and we would expect that already very local information is sufficient to separate the fields. Moreover, out of all one-unit intervals, the shapes of the three covariance functions differ the most from each other in the interval from 0 to 1 and we may reasonably assume that either $B(1)$ or $\{B(1), R(1,2)\}$ will be the most efficient choice.

The mean values of $n(p-1)\mathrm{MDI}(\widehat{\Gamma})^2$ over the 2000 replications are shown as the solid lines in Figure 3 with the dashed lines representing the limiting values towards which the means of the quantity are expected to converge (see Propositions 2 and 4). As evidenced in Figure 3, this is indeed what occurs, with the sawtooth form most likely being an artefact of our highly geometric choice of a location pattern. Unsurprisingly, $B(1)$ is the most efficient choice of kernel, in both the limit and for finite samples, and also surpasses the joint diagonalization of $B(1)$ and $R(1,2)$ in efficiency. We also note that the joint use of the two kernels is more effective than using only the ring kernel $R(1,2)$, which does not utilize the interval $(0,1)$ where the largest differences in the shapes of the covariance functions lie.

The previous investigation and Figure 3 used only the expected value of the limiting

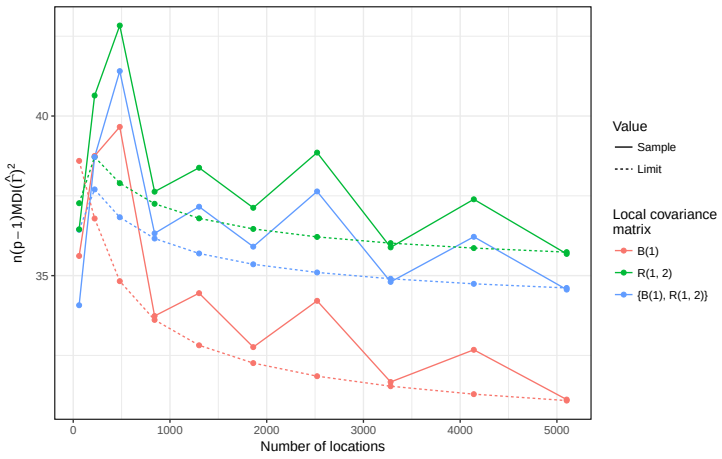

Figure 3: The mean values of $n(p-1)\mathrm{MDI}(\widehat{\Gamma})^2$ in the first simulation. The dashed lines correspond to the limiting values of the same quantity.

distribution. In Figure 4 we have plotted the estimated densities of $n(p-1)\mathrm{MDI}(\widehat{\Gamma})^2$ (black lines) against the corresponding limiting densities (red lines) for all kernels and a few selected sample sizes. The density functions of the limiting distributions are estimated from a sample of 100,000 random variables drawn from the corresponding distributions. Overall, the two densities fit each other rather well for all kernels and especially for $n = 221$ (which corresponds to a diamond of radius $m = 10$). Apart from the peak, the fit is particularly good. This shows that the limiting distribution of $n(p-1)\mathrm{MDI}(\widehat{\Gamma})^2$ is a good approximation of the true distribution already for very small sample sizes.

5.2 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 comparison is made on the basis of "asymptotic efficiency", the expected value of the limiting distribution of $n(p-1)\mathrm{MDI}(\widehat{\Gamma})^2$ using the asymptotic distribution given in Propositions 2 and 4, 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 respectives shape parameters $\kappa = 2, 1, 0.25$ and a range parameter $\phi \in \{1.0, 1.1, 1.2, \ldots, 30.0\}$. The three covariance functions are shown for $\phi = 1$ in the right panel of Figure 1. The random field is observed at three different point patterns: diamond-shaped, rectangular and random (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 while 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 two neighbouring locations is one unit and examples of the two pattern types are shown in Figure 2 with a radius $m = 10$.

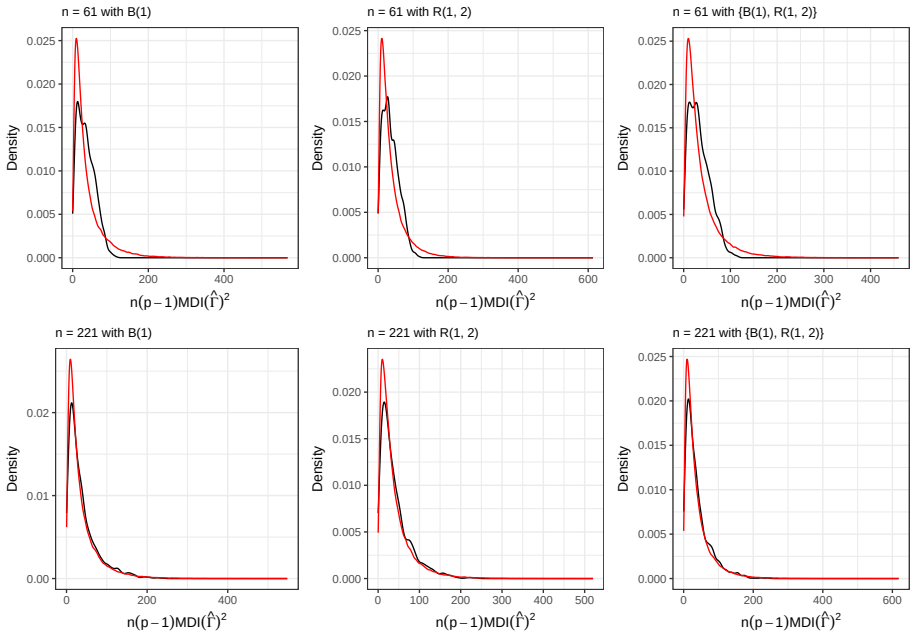

Figure 4: The density of $n(p-1)\mathrm{MDI}(\widehat{\Gamma})^2$ over 2000 replications (in black) against the density of its limiting distribution (in red) for different combinations of sample size and local covariance matrices.

A rectangular pattern with "radius" m is defined to have the width $2m+1$ and the height $m+1$. The random point pattern is generated simply by simulating $n = 1861$ points uniformly in the rectangle $(-30, 30) \times (-30, 30)$. These particular point patterns were selected for illustration as their sizes can be easily varied to incorporate different sample sizes. We consider a total of eight different local covariance matrices, $B(r), R(r-1, r)$ for $r = 1, 3, 5$, and the joint diagonalizations of the previous 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 Figure 5 where the two joint diagonalizations are denoted by having value "J" as the parameter r. Recall that the lower the value on the y-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 the location pattern does not have a large effect on the results. In all the patterns the local covariance matrices with either $r = 1$ or $r = 3$ are the best choices for small values of the range ϕ but they quickly deteriorate as ϕ increases. The opposite happens for the local covariance matrices with $r = 5$; they are among the worst for small ϕ and relatively improve with increasing ϕ. The joint diagonalization-based choices fall somewhere in between and are never the best nor the worst choice. However, they yield performance very close to the best choice in the right end of the range-scale and are close to the optimal ones in the left end. Thus, their use could be justified in practice as the "safe choice". Comparing the two types

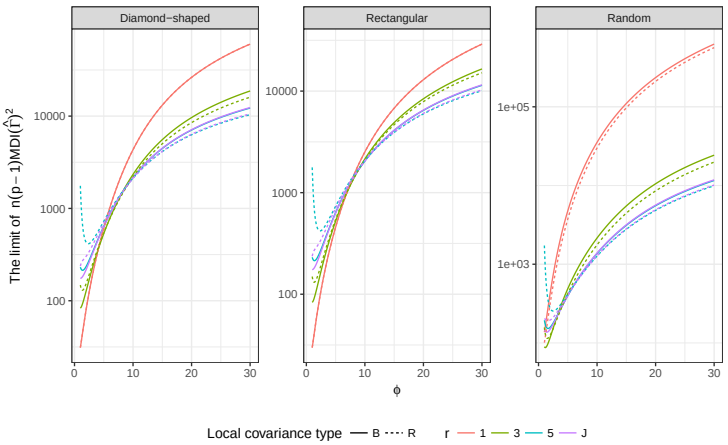

Figure 5: The limiting values of $n(p-1)\mathrm{MDI}(\widehat{\Gamma})^2$ as a function of the range of the underlying Matérn random fields for the different choices of local covariance matrices in the second simulation. The y-axis has a logarithmic scale.

of local covariance matrices, balls and rings, we observe that in the majority of cases the rings prove superior to the balls.

5.3 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 the shape parameters $\kappa = 6, 1, 0.25$ and the range parameter $\phi = 20$ (in kilometers), see the left panel of Figure 1. We consider two different, fixed location patterns fitted inside the map of Finland, see Figure 6. The first location pattern has the locations drawn uniformly from the map and the second location 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 the respective radii of 10, 20, and 30 kilometers 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 MDI of the estimates obtained with the local covariance matrix kernels $B(r), R(r-10, r), G(r)$, where $r = 10, 20, 30, 100$, 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 $\{G(10), G(20), G(30), G(100\}$ adding up to a total of 15 estimators. The Gaussian kernel is parametrized 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, making $G(r)$ have 90% of its total mass in the radius r ball around its center. Thus $G(r)$ can be considered as a smooth approximation of $B(r)$. The larger radius kernels $B(100)$, $R(90, 100)$, $G(100)$ are included in the simulation to investigate what happens when we overestimate the dependency radius.

Figure 6: The two fixed location patterns in the map of Finland (left: uniform map, right: skew map).

The mean minimum distance index values for the 15 estimators are plotted in Figure 7 and show that for both maps and all local covariance types, increasing the radius yields more accurate separation results all the way up to $r = 30$, whereas for $r = 100$ the results again worsen. This observation shows that when using a single local covariance matrix the choice of the type and the radius are especially important, most likely requiring some expert knowledge on the study. However, this problem is completely averted when we use the joint diagonalization of several matrices. For both maps and all local covariance types the joint diagonalization produces results very comparable to the best individual matrices, even though the joint diagonalizations also include the "bad choices", $r = 10, 20, 100$. We also observed similar behaviour in the first and second simulation studies where, in the absence of knowledge on the optimal choice, the joint diagonalization proved a middle ground between the best and worst choices. Thus, we recommend the use of the joint diagonalization of scatter matrices with a sufficiently large variation of radii for the kernels.

Finally, a comparison between 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 (the minimum distance index values are on average higher). This is explained by the large number of isolated points which 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 which is

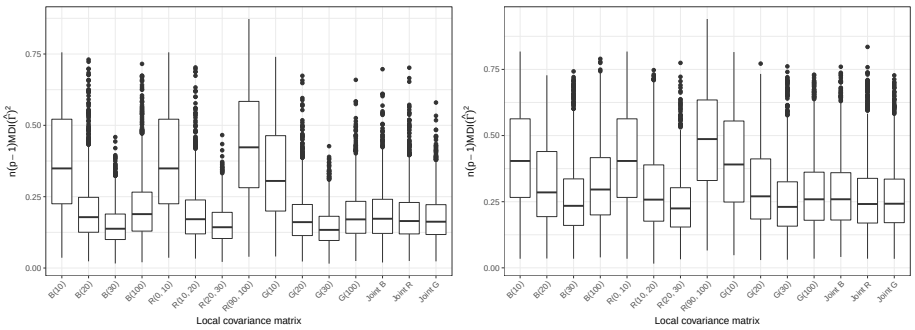

Figure 7: The results of the efficiency study. On the left the results for the uniform sampling design and on the right for the skew design.

Table 1: Absolute values of the maximal correlations of different estimators with respect to the standard. All estimators used the empirical covariance matrix. The distances for the scatters are given in kilometers. Est, estimator; C, component.

Est	Scatters	C1	C2	C3	C4	C5	C6
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

available in the R package StatDa (Filzmoser, 2015) and described in Reimann et al. (2008), for example. The data consists of 594 samples of terrestrial moss collected at different sites in north Europe on the borders of Norway, Finland and Russia. The amount of 31 chemical elements found in the moss samples was already used as a spatial blind source separation example in Nordhausen et al. (2015) where the covariance matrix and $B(50)$ were simultaneously diagonalized. The radius of 50 kilometers was carefully chosen by a subject expert in that analysis and considered best compared to several other radii not mentioned there. The analysis found six meaningful components, which could be used to distinguish underlying natural geological patterns from environmental pollution patterns. For the analysis we follow the same data preparation steps as in Nordhausen et al. (2015) and then compute several competing spatial blind source separation estimates. Using these methods, we identify the six components with the highest correlations (in absolute values) with respect to the six main components identified in Nordhausen et al. (2015). The different competing estimators and their components' correlations to the components of the standard are given in Table 1. The table shows that when using only two scatters (Estimators 1, 2, 3), some components cannot be easily found. However, when jointly diagonalizing more than two scatters

the results are more stable and less dependent on the chosen distances of the scatters (Estimators 4, 5).

A Appendix

Proposition 1, (i) and (ii), correspond to Propositions A.1 and A.2 in the Appendix. Proposition 2 corresponds to Proposition A.4 in the Appendix. Finally, Propositions 3 and 4 follow directly from Proposition A.8 in the Appendix.

A.1 General results

Recall that $d \in \mathbb{N}$ and $p \in \mathbb{N}$ are fixed. $Z_1, \ldots, Z_p$ are p independent stationary Gaussian processes on $\mathbb{R}^d$ with zero mean functions, unit variances and covariance functions $K_1, \ldots, K_p$. We have $Z = (Z_1, \ldots, Z_p)^{\mathrm{T}}$ and $X = (X_1, \ldots, X_p)^{\mathrm{T}} = \Omega Z$ with Ω a fixed invertible $p \times p$ matrix.

Let $s_1, \ldots, s_n$ be the n observation points in $\mathcal{S}^d$ and let f be a kernel function from $\mathbb{R}^d$ into $\mathbb{R}$. We recall

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

and

$$M(f) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} f(s_i - s_j) \Omega D(s_i, s_j) \Omega^{\mathrm{T}}$$

where $D(s_i, s_j)$ is the $p \times p$ diagonal matrix defined by

$$D(s_i, s_j)_{k,k} = K_k(s_i - s_j).$$

Let $|x| = \max_{i=1,\ldots,m} |s_i|$ be the sup norm for $x \in \mathbb{R}^m$. Since this norm is equivalent to the Euclidian norm, Assumptions 5 to 7 are equivalent to the following conditions.

Condition A.1. *There exists a fixed $\Delta > 0$ so that for all $n \in \mathbb{N}$ and for all $a \neq b$, $a, b \in \{1, \ldots, n\}$, $|s_a - s_b| \geq \Delta$.*

Condition A.2. *There exist a fixed $A < +\infty$ and $\alpha > 0$ so that for all $s \in \mathbb{R}^d$ and for all $k = 1, \ldots, p$,*

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

Condition A.3. *There exist a fixed $A < +\infty$ and $\alpha > 0$ so that for all $s \in \mathbb{R}^d$ and for all $k = 1, \ldots, p$,*

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

For a matrix M, denote by $M_{i,j}$ the element from the i-th row and the j-th column of M. For a vector V_n or a matrix M_n, denote by $[V_n]_i$ the i-th element of V_n and by $[M_n]_{i,j}$ the element from the i-th row and the j-th column of M_n. The singular values of a $n \times n$ matrix M are denoted by $\rho_1(M) \geq \cdots \geq \rho_n(M) \geq 0$ and, in the case when M is symmetric, the eigenvalues are denoted by $\lambda_1(M) \geq \cdots \geq \lambda_n(M)$. The spectral norm is given by $\rho_1(M)$ and $\|M\|_F^2 = \sum_{i,j} |M_{i,j}|^2$ denotes the Frobenius norm. For a sequence of random variables X_n, we write $X_n = o_p(1)$ when X_n converges to 0 in probability as $n \to \infty$ and we write $X_n = O_p(1)$ when X_n is bounded in probability as $n \to \infty$. Let $e_i(k)$ be the i-th base column vector of $\mathbb{R}^k$. Let y be the $np \times 1$ vector defined by $y_{(i-1)p+j} = X_j(s_i)$, for $i = 1, \ldots, n$, $j = 1, \ldots, p$.

Lemma A.1. *Under Conditions A.1 and A.2, there exists a finite constant $C < +\infty$ so that for all $n \in \mathbb{N}$,*

$$\lambda_1\{\mathrm{cov}(y)\} \leq C.$$

Proof. Let ψ_a denote the ath row of Ω. We have

$$\begin{aligned} |\mathrm{cov}\{X_a(s_i), X_b(s_j)\}| &= |\mathrm{cov}\{\psi_a Z(s_i), \psi_b Z(s_j)\}| \\ &= |\psi_a D(s_i, s_j) \psi_b^{\mathrm{T}}| \\ &\leq \max_{a,b=1,\ldots,p} [\Omega_{a,b}]^2 \frac{A}{1 + |s_i - s_j|^{d+\alpha}} \end{aligned} \tag{2}$$

from Condition A.2. Note also that $\lambda_1\{\mathrm{cov}(y)\} = \lambda_1\{\mathrm{cov}(\tilde{y})\}$ where $\tilde{y}$ is the $np \times 1$ vector defined by $\tilde{y}_{(j-1)n+i} = X_j(s_i)$ for $i = 1, \ldots, n$, $j = 1, \ldots, p$. Hence, the lemma is a direct consequence of Lemma 6 in Furrer et al. (2016). □

The next theorem provides a general multivariate central limit theorem for quadratic forms of Gaussian vectors. It extends standard central limit theorems in spatial statistics (see, e.g., Bachoc (2014) or Istas and Lang (1997)) by allowing cases where the sequence of covariance matrices is non-converging or asymptotically singular. The full proof is given for self-consistency, although some of the arguments have appeared previously.

Theorem A.1. *Let (y_n) be a sequence of n-dimensional centered Gaussian vectors. Let R_n be the covariance matrix of y_n. Assume that for all n, $\lambda_1(R_n) \leq A$ where A is a fixed finite constant. Let $k \in \mathbb{N}$ be fixed and let $(M_{1,n}), \ldots, (M_{k,n})$ be k sequences of deterministic $n \times n$ symmetric matrices. Assume that for $i = 1, \ldots, k$, for $n \in \mathbb{N}$, $\rho_1(M_{i,n}) \leq A$. Let Σ_n be the $k \times k$ matrix defined for $1 \leq i, j \leq k$, by*

$$[\Sigma_n]_{i,j} = 2n^{-1}\mathrm{Tr}\left(R_n M_{i,n} R_n M_{j,n}\right).$$

Let r_n be the k-dimensional vector defined for $i = 1, \ldots, k$, by

$$[r_n]_i = \mathrm{Tr}\left(n^{-1} R_n M_{i,n}\right).$$

Let V_n be the $k \times 1$ vector defined for $1 \leq i \leq k$, by

$$[V_n]_i = n^{-1} y_n^{\mathrm{T}} M_{i,n} y_n.$$

Let Q_n be the probability measure of $n^{1/2}(V_n - r_n)$ on $\mathbb{R}^k$. Let $\mathcal{N}(0, \Sigma_n)$ be the Gaussian distribution on $\mathbb{R}^k$ with mean vector 0 and covariance matrix Σ_n. Let d_w denote a metric generating the topology of weak convergence on the set of Borel probability measures on $\mathbb{R}^k$; for specific examples see the discussion in Dudley (2002) p. 393. Then we have, for $n \to \infty$,

$$d_w\{Q_n, \mathcal{N}(0, \Sigma_n)\} \to 0.$$

Proof. Assume that $d_w\{Q_n, \mathcal{N}(0, \Sigma_n)\} \not\to 0$ when $n \to \infty$. Then there exists $\epsilon > 0$ fixed and a subsequence n_m so that $d_w\{Q_{n_m}, \mathcal{N}(0, \Sigma_{n_m})\} \geq \epsilon$. For $a_1, \ldots, a_k \in \mathbb{R}$, let $S_{n_m} = R_{n_m}^{1/2}(\sum_{i=1}^k a_i M_{i,n_m}) R_{n_m}^{1/2}$. We have

$$\sum_{i,j=1}^k a_i a_j [\Sigma_{n_m}]_{i,j} = 2\mathrm{Tr}\left(S_{n_m}^2\right)/n_m.$$

Hence, we see that Σ_{n_m} is a non-negative matrix, and, from the assumptions on (R_{n_m}) and (M_{i,n_m}), that $[\Sigma_{n_m}]_{i,i} \leq 2A^4$. Also, $|[r_n]_i| = |n^{-1}\mathrm{Tr}(R_n^{1/2} M_{i,n} R_n^{1/2})| \leq A^2$. Hence, by compacity, and up to extracting a further subsequence, we can assume that $r_{n_m} \to r$ and $\Sigma_{n_m} \to \Sigma$ when $n_m \to \infty$. One can show simply that $d_w\{\mathcal{N}(0, \Sigma_{n_m}), \mathcal{N}(0, \Sigma)\} \to 0$ when $n_m \to \infty$. Hence, when $n_m \to \infty$,

$$\limsup d_w\{Q_{n_m}, \mathcal{N}(0, \Sigma)\} \geq \epsilon. \tag{3}$$

Let $a_1, \ldots, a_k \in \mathbb{R}$ be fixed. Then, let $z_{n_m} = R_{n_m}^{-1/2} y_{n_m}$, $S_{n_m} = P_{n_m} D_{n_m} P_{n_m}^{\mathrm{T}}$ with diagonal D_{n_m} and $P_{n_m} P_{n_m}^{\mathrm{T}} = I_{n_m}$ and let $\xi_{n_m} = P_{n_m}^{\mathrm{T}} z_{n_m}$. Observe that z_{n_m} and ξ_{n_m} follow the $\mathcal{N}(0, I_{n_m})$ distribution. We have

$$\begin{aligned}
\sum_{i=1}^k a_i [V_n]_i &= n_m^{-1} y_{n_m}^{\mathrm{T}} (\sum_{i=1}^k a_i M_{i,n_m}) y_{n_m} \\
&= n_m^{-1} z_{n_m}^{\mathrm{T}} S_{n_m} z_{n_m} \\
&= n_m^{-1} \sum_{a=1}^{n_m} [\xi_{n_m}]_a^2 \lambda_a\left(S_{n_m}\right).
\end{aligned}$$

Hence

$$\begin{aligned}
W_{n_m} &:= {n_m}^{1/2} (\sum_{i=1}^k a_i [V_{n_m}]_i - \sum_{i=1}^k a_i [r_{n_m}]_i) \\
&= {n_m}^{-1/2} \sum_{\ell=1}^{n_m} ([\xi_{n_m}]_\ell^2 - 1) \lambda_\ell\left(S_{n_m}\right).
\end{aligned}$$

If $\sum_{i,j=1}^k a_i a_j \Sigma_{i,j} = 0$ when $n_m \to \infty$, then $\sum_{i,j=1}^k a_i a_j [\Sigma_{n_m}]_{i,j} \to 0$. Hence, $2n_m^{-1}\mathrm{Tr}(S_{n_m}^2) \to 0$ and so $\mathrm{var}(W_{n_m}) \to 0$. Hence $W_{n_m} \to \mathcal{N}(0,0) = \mathcal{N}(0, \sum_{i,j=1}^k a_i a_j \Sigma_{i,j})$ in distribution when $n_m \to \infty$.

Now, if $\sum_{i,j=1}^{k} a_i a_j \Sigma_{i,j} > 0$, one can show from the Lindeberg-Feller central limit theorem that when $n_m \to \infty$, $W_{n_m} \to \mathcal{N}(0, \sum_{i,j=1}^{k} a_i a_j \Sigma_{i,j})$ in distribution (see also Lemma 2 in Istas and Lang (1997)).

Hence, since both of the above-considered convergences in distribution hold for any $a_1, \ldots, a_k$, we have, by Cramér-Wold theorem, that when $n_m \to \infty$, ${n_m}^{1/2}(V_{n_m} - r_{n_m}) \to \mathcal{N}(0, \Sigma)$ in distribution. This is in contradiction with (3). Hence when $n \to \infty$

$$d_w\{Q_n, \mathcal{N}(0, \Sigma_n)\} \to 0.$$

□

A.2 Asymptotics when diagonalizing two matrices

The next proposition gives the consistency of $\widehat{M}(f)$.

Proposition A.1. *Let Conditions A.1 and A.2 hold and let $f : \mathbb{R}^d \to \mathbb{R}$ satisfy Condition A.3. Then as $n \to \infty$, $\widehat{M}(f) - M(f) \to 0$ in probability.*

Proof. Clearly $E\{\widehat{M}(f)\} = M(f)$. Let $k, l \in \{1, \ldots, p\}$ be fixed. In order to prove the proposition, it is sufficient to show that when $n \to \infty$, $\mathrm{var}\{\widehat{M}(f)_{k,l}\} \to 0$.

We have,

$$\begin{aligned}
\widehat{M}(f)_{k,l} &= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} f(s_i - s_j) e_k(p)^{\mathrm{T}} X(s_i) X(s_j)^{\mathrm{T}} e_l(p) \\
&= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} f(s_i - s_j) X(s_i)^{\mathrm{T}} e_k(p) e_l(p)^{\mathrm{T}} X(s_j) \\
&= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} f(s_i - s_j) X(s_i)^{\mathrm{T}} [(1/2)\{e_k(p) e_l(p)^{\mathrm{T}} + e_l(p) e_k(p)^{\mathrm{T}}\}] X(s_j).
\end{aligned}$$

Let $T_{k,l}(f)$ be the $np \times np$ matrix, that we see as a block matrix composed of n^2 blocks of sizes p^2, and with block i, j equal to $f(s_i - s_j)(1/2)\{e_k(p) e_l(p)^{\mathrm{T}} + e_l(p) e_k(p)^{\mathrm{T}}\}$. With this notation,

$$\widehat{M}(f)_{k,l} = n^{-1} y^{\mathrm{T}} T_{k,l}(f) y.$$

The largest singular value of $T_{k,l}(f)$ is bounded as $n \to \infty$. Indeed, from Gershgorin's circle theorem, $\rho_1\{T_{k,l}(f)\}$ is no larger than $\max_{i=1,\ldots,np} \sum_{j=1}^{np} |T_{k,l}(f)_{i,j}|$. This maximum is no larger than $\max_{i=1,\ldots,n} \sum_{j=1}^{n} |f(s_i - s_j)|$. This last quantity is bounded as $n \to \infty$ from Condition A.3 and Lemma 4 in Furrer et al. (2016).

Hence $\rho_1\{T_{k,l}(f)\}$ is bounded by a constant $B < +\infty$. Thus

$$\mathrm{var}\{\widehat{M}(f)_{k,l}\} = 2n^{-2} \mathrm{Tr}\{\mathrm{cov}(y) T_{k,l}(f) \mathrm{cov}(y) T_{k,l}(f)\} \leq 2pn^{-1} B^2 C^2,$$

with $\lambda_1\{\mathrm{cov}(y)\} \leq C$ from Lemma A.1. □

The next proposition is a corollary of Theorem A.1 and gives the asymptotic normality of $\widehat{M}(f)$.

Proposition A.2. *Let, for $k, l = 1, \ldots, p$ and $f : \mathbb{R}^d \to \mathbb{R}$, $T_{k,l}(f)$ be defined as in the proof of Proposition A.1. Let $R = \mathrm{cov}(y)$ and let $\Sigma(f)$ be the $p^2 \times p^2$ matrix defined by, for $i = (s-1)p + t$ and $j = (u-1)p + v$, with $s, t, u, v \in \{1, \ldots, p\}$,*

$$\Sigma(f)_{i,j} = 2n^{-1}\mathrm{Tr}\left\{RT(f)_{s,t}RT(f)_{u,v}\right\}.$$

Define, for $g : \mathbb{R}^d \to \mathbb{R}$, $\Sigma(f, g)$ as the $p^2 \times p^2$ matrix defined for $i = (s-1)p + t$ and $j = (u-1)p + v$, with $s, t, u, v \in \{1, \ldots, p\}$ by

$$\Sigma(f, g)_{i,j} = 2n^{-1}\mathrm{Tr}\left\{RT(f)_{s,t}RT(g)_{u,v}\right\}.$$

Let

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

Assume that Conditions A.1 and A.2 hold. Let $f_1, f_2 : \mathbb{R}^d \to \mathbb{R}$ satisfy Condition A.3. Let for $r = 1, 2$, $W(f_r)$ be the vector of size $p^2 \times 1$, defined for $i = (a-1)p + b$, $a, b \in \{1, \ldots, p\}$, by $W(f_r)_i = n^{1/2}\{\widehat{M}(f_r)_{a,b} - M(f_r)_{a,b}\}$.

Let Q_n be the distribution of $\{W(f_1)^{\mathrm{T}}, W(f_2)^{\mathrm{T}}\}^{\mathrm{T}}$. Then as $n \to \infty$

$$d_w[Q_n, \mathcal{N}\{0, V(f_1, f_2)\}] \to 0.$$

Furthermore $\lambda_1\{V(f_1, f_2)\}$ is bounded as $n \to \infty$.

Proposition 1(ii) is a direct corollary of Proposition A.2 with $f_1 = f$ and $f_2 = f_0$. Moreover, Proposition A.2 gives details concerning the matrix $V(f_1, f_2)$.

Proof. Let $a, b \in \{1, \ldots, p\}$. We have seen in the proof of Proposition A.1 that for $r = 1, 2$,

$$\widehat{M}(f_r)_{a,b} = n^{-1}y^{\mathrm{T}}T_{a,b}(f_r)y,$$

$$E(\widehat{M}(f_r)_{a,b}) = M(f_r)_{a,b} = n^{-1}\mathrm{Tr}\{RT_{a,b}(f_r)\}.$$

Furthermore, we have, from Lemma A.1 and the proof of Proposition A.1, that $\lambda_1(R)$ and $\rho_1\{T_{a,b}(f_r)\}$ are bounded as $n \to \infty$. Hence, the proposition is a consequence of Theorem A.1. Finally, $\lambda_1\{V(f_1, f_2)\}$ is bounded as $n \to \infty$ because each component of $V(f_1, f_2)$ is bounded as $n \to \infty$. □

Recall that $\mathrm{vect}(M) = (l_1^{\mathrm{T}}, \ldots, l_k^{\mathrm{T}})^{\mathrm{T}}$ where $l_1^{\mathrm{T}}, \ldots, l_k^{\mathrm{T}}$ are the k rows of a matrix M. We let $\mathcal{D}(k) = \{1 + (i-1)(k+1); i = 1, \ldots, k\}$. Also $\{\mathrm{vect}(M)_i; i \in \mathcal{D}(k)\} = \{M_{i,i}; i = 1, \ldots, k\}$. Let $\bar{\mathcal{D}}_k = \{1, \ldots, k^2\} \backslash \mathcal{D}_k$. Also $\{\mathrm{vect}(M)_i; i \in \bar{\mathcal{D}}(k)\} = \{M_{i,j}; i, j = 1, \ldots, k, i \neq j\}$. For $a \in \{1, \ldots, k^2\}$, let $I_k(a)$ and $J_k(a)$ be the unique $i, j \in \{1, \ldots, k\}$ so that $a = k(i-1) + j$. For $i \in \{1, \ldots, k\}$, let $d_k(i) = 1 + (i-1)(k+1)$ and note that $\{\mathrm{vect}(M)_{d_k(i)}; i = 1, \ldots, k\} = \{M_{i,i}; i = 1, \ldots, k\}$. For a matrix M of size $k \times k$, recall that $\mathrm{diag}(M) = (M_{1,1}, \ldots, M_{k,k})^{\mathrm{T}}$.

Let $D_0 = D(0, 0) = I_p$ and let

$$\widehat{D}_0 = n^{-1}\sum_{i=1}^{n} Z(s_i)Z(s_i)^{\mathrm{T}}.$$

For $f : \mathbb{R}^d \to \mathbb{R}$, let

$$D(f) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} f(s_i - s_j) D(s_i, s_j)$$

and

$$\widehat{D}(f) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} f(s_i - s_j) Z(s_i) Z(s_j)^{\mathrm{T}}.$$

Let $f : \mathbb{R}^d \to \mathbb{R}$ satisfy Condition A.3. Let $\lambda_1 > \ldots > \lambda_p$ be the diagonal elements of $D(f)$, that we assume to be two-by-two distinct for all $n \in \mathbb{N}$.

Proposition A.3. *Let A be the $p^2 \times p^2$ matrix defined by*

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

Let B be the $p^2 \times p^2$ matrix defined by

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

Let C be the $p \times p^2$ matrix defined by

$$C_{i,j} = \begin{cases} -\lambda_i & \text{for } j = d_k(i), \\ 0 & \text{otherwise}. \end{cases}$$

Let D be the $p \times p^2$ matrix defined by

$$D_{i,j} = \begin{cases} 1 & \text{for } j = d_k(i), \\ 0 & \text{otherwise}. \end{cases}$$

Assume that when $n \to \infty$, $\liminf(\min_{i \neq j} |\lambda_i - \lambda_j|) > 0$. Then, with probability going to one as $n \to \infty$, there exist two sequences of random matrices $\Gamma\{\widehat{D}_0, \widehat{D}(f)\}$ and $\Lambda\{\widehat{D}_0, \widehat{D}(f)\}$, satisfying Definition 2 with $\widehat{M}(f_0)$, $\widehat{M}(f)$ replaced by $\widehat{D}_0$, $\widehat{D}(f)$ and such that, as $n \to \infty$,

$$\begin{pmatrix} n^{1/2}(\mathrm{vect}[\Gamma\{\widehat{D}_0, \widehat{D}(f)\} - I_p]) \\ n^{1/2}(\mathrm{diag}[\Lambda\{\widehat{D}_0, \widehat{D}(f)\} - D(f)]) \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} n^{1/2}\{\mathrm{vect}(\widehat{D}_0 - D_0)\} \\ n^{1/2}[\mathrm{vect}\{\widehat{D}(f) - D(f)\}] \end{pmatrix} + o_p(1).$$

Proof. From Proposition A.1, with probability going to one, the eigenvalues of $\widehat{D}(f)$ are two-by-two distinct. In the rest of the proof, we set ourselves on the event when this is the case. Then, let $\widehat{\Gamma} = \Gamma\{\widehat{D}_0, \widehat{D}(f)\}$ and $\widehat{\Lambda} = \Lambda\{\widehat{D}_0, \widehat{D}(f)\}$ for which

$$\widehat{\Lambda}_{1,1} \geq \cdots \geq \widehat{\Lambda}_{p,p} \quad \text{and} \quad \sum_{j=1}^{p} \widehat{\Gamma}_{i,j} \geq 0, \quad \text{for } i = 1, \ldots, p. \tag{4}$$

Let

$$T_1 = \begin{pmatrix} n^{1/2}\mathrm{vect}[\Gamma\{\widehat{D}_0, \widehat{D}(f)\} - I_p] \\ n^{1/2}\mathrm{diag}[\Lambda\{\widehat{D}_0, \widehat{D}(f)\} - D(f)] \end{pmatrix}$$

and

$$T_2 = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} n^{1/2}\{\mathrm{vect}(\widehat{D}_0 - D_0)\} \\ n^{1/2}[\mathrm{vect}\{\widehat{D}(f) - D(f)\}] \end{pmatrix}.$$

Assume that $T_1 - T_2 \not\to 0$ in probability when $n \to \infty$. Then there exist $\epsilon > 0$ and a subsequence $n_m \to \infty$ so that along n_m

$$P(|T_1 - T_2| \geq \epsilon) \geq \epsilon. \tag{5}$$

One can show, as for the proof of Proposition A.1, that $\limsup \lambda_1\{D(f)\} < +\infty$ when $n \to \infty$. Hence, up to extracting a further subsequence, we can assume that when $n_m \to \infty$, $D(f) \to D_\infty(f)$, where $D_\infty(f)$ has two-by-two distinct, decreasing, eigenvalues.

From Lemma 4.3 in Sun and Sun (2002), since $D_0^{-1/2} D_\infty(f) D_0^{-1/2} = D_\infty(f)$ is diagonal, there exists a sequence of random orthogonal matrix U_n so that $U_{n_m}^{\mathrm{T}} \widehat{D}_0^{-1/2} \widehat{D}(f) \widehat{D}_0^{-1/2} U_{n_m} = \Lambda_{n_m}$ is diagonal and goes to $D_\infty(f)$ in probability and so that $U_{n_m} \to I_p$ in probability when $n_m \to \infty$. Hence, $U_{n_m} \widehat{D}_0^{-1/2}, \Lambda_{n_m}$ satisfies (4) asymptotically, so that it must coincide with $\widehat{\Gamma}$ and $\widehat{\Lambda}$ asymptotically . Indeed, any other matrix obtained by permuting or changing the signs of the rows of $\widehat{\Gamma}$ and $\widehat{\Lambda}$ cannot satisfy (4) asymptotically. Hence we have finally obtained $\widehat{\Gamma} \to I_p$ and $|\widehat{\Lambda} - D(f)| \to 0$ in probability when $n_m \to \infty$.

The rest of the proof is similar to those given in Ilmonen et al. (2010a) and Miettinen et al. (2012). By definition of $\widehat{\Gamma}$ and $\widehat{\Lambda}$, we have

$$\widehat{\Gamma}\widehat{D}_0\widehat{\Gamma}^{\mathrm{T}} = I_p \text{ and } \widehat{\Gamma}\widehat{D}(f)\widehat{\Gamma}^{\mathrm{T}} = \widehat{\Lambda}.$$

Hence

$$\begin{gathered}(\widehat{\Gamma} - I_p)\widehat{D}_0\widehat{\Gamma}^{\mathrm{T}} + (\widehat{D}_0 - I_p)\widehat{\Gamma}^{\mathrm{T}} + (\widehat{\Gamma} - I_p)^{\mathrm{T}} = 0 \text{ and} \\ (\widehat{\Gamma} - I_p)\widehat{D}(f)\widehat{\Gamma}^{\mathrm{T}} + \{\widehat{D}(f) - D(f)\}\widehat{\Gamma}^{\mathrm{T}} + D(f)(\widehat{\Gamma} - I_p)^{\mathrm{T}} = \widehat{\Lambda} - D(f).\end{gathered}$$

Also, from Proposition A.2 (taken in the case $\Omega = I_p$), we have ${n_m}^{1/2}(\widehat{D}_0 - I_p) = O_p(1)$ and ${n_m}^{1/2}\{\widehat{D}(f) - D(f)\} = O_p(1)$. Thus, we get

$$\begin{aligned} {n_m}^{1/2}(\widehat{D}_0 - I_p) &= -{n_m}^{1/2}(\widehat{\Gamma} - I_p) - {n_m}^{1/2}(\widehat{\Gamma} - I_p)^{\mathrm{T}} + o_p(1) \text{ and} \\ {n_m}^{1/2}\{\widehat{D}(f) - D(f)\} &= -{n_m}^{1/2}(\widehat{\Gamma} - I_p)D(f) - {n_m}^{1/2}D(f)(\widehat{\Gamma} - I_p)^{\mathrm{T}} + {n_m}^{1/2}\{\widehat{\Lambda} - D(f)\} + o_p(1). \end{aligned}$$

This then yields

$$\begin{gathered} {n_m}^{1/2}(\widehat{\Gamma}_{ii} - 1) = -\tfrac{1}{2}{n_m}^{1/2}(\widehat{D}_{0,i,i} - 1) + o_p(1) \\ (\lambda_i - \lambda_j){n_m}^{1/2}\widehat{\Gamma}_{i,j} = {n_m}^{1/2}\widehat{D}(f)_{i,j} - \lambda_i {n_m}^{1/2}\widehat{D}_{0,i,j} + o_p(1), \;\; i \neq j, \text{ and} \\ {n_m}^{1/2}(\widehat{\Lambda}_{i,i} - \lambda_i) = {n_m}^{1/2}\{\widehat{D}(f)_{i,i} - \lambda_i\} - \lambda_i {n_m}^{1/2}(\widehat{D}_{0,i,i} - 1) + o_p(1). \end{gathered}$$

This is in contradiction with (5), by definition of A, B, C and D. Hence the proof is finished. □

Let now $\widehat{M}_0 = \widehat{M}(f_0)$ where $f_0(x) = \mathbf{1}_{\{x=0\}}$. Let $\Gamma\{\widehat{M}_0, \widehat{M}(f)\} = \widehat{\Gamma}(f)$ and $\Lambda\{\widehat{M}_0, \widehat{M}(f)\} = \widehat{\Lambda}(f)$ with the notation of Definition 2.

Proposition A.4. *Assume the same conditions as in Proposition A.3. Let*

$$E = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

from Proposition A.3. Let $M_{\Omega^{-1}}$ be the $p^2 \times p^2$ matrix defined by

$$[M_{\Omega^{-1}}]_{a,b} = \begin{cases} [\Omega^{-1}]_{J_p(b),J_p(a)} & \text{if } I_p(a) = I_p(b), \\ 0 & \text{if } I_p(a) \neq I_p(b). \end{cases}$$

Let $\bar{M}_{\Omega^{-1}}$ be the matrix of size $(p^2+p) \times (p^2+p)$ defined by

$$\bar{M}_{\Omega^{-1}} = \begin{pmatrix} M_{\Omega^{-1}} & 0 \\ 0 & I_p \end{pmatrix}.$$

Let

$$X_n = n^{1/2} \begin{pmatrix} \mathrm{vect}[\Gamma\{\widehat{M}_0, \widehat{M}(f)\} - \Omega^{-1}] \\ \mathrm{diag}[\Lambda\{\widehat{M}_0, \widehat{M}(f)\} - D(f)] \end{pmatrix}.$$

Let Q_n be the distribution of X_n. Let $\tilde{V}(f)$ be as $V(f_0, f)$ in Proposition A.2 but where R is replaced by $\mathrm{cov}(z)$ where z is the $np \times 1$ vector defined for $i = 1, \ldots, n$, $j = 1, \ldots, p$, by $z_{(i-1)p+j} = Z_j(s_i)$. Let

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

Then, we can choose $\Gamma\{\widehat{M}_0, \widehat{M}(f)\}, \Lambda\{\widehat{M}_0, \widehat{M}(f)\}$ so that when $n \to \infty$,

$$d_w\{Q_n, \mathcal{N}(0, F)\} \to 0.$$

Proof. From Proposition A.3, because $\widehat{M}(f) = \Omega \widehat{D}(f) \Omega^{\mathrm{T}}$ for any $f : \mathbb{R}^d \to \mathbb{R}$, with probability going to one, $\Gamma\{\widehat{M}_0, \widehat{M}(f)\}, \Lambda\{\widehat{M}_0, \widehat{M}(f)\}$ exist and can be chosen so that

$$n^{1/2} \mathrm{vect}[\Gamma\{\widehat{M}_0, \widehat{M}(f)\} - \Omega^{-1}] = n^{1/2} \mathrm{vect}([\Gamma\{\widehat{D}_0, \widehat{D}(f)\} - I_p]\Omega^{-1}),$$

$$n^{1/2} \operatorname{diag}[\Lambda\{\widehat{M}_0, \widehat{M}(f)\} - D(f)] = n^{1/2} \operatorname{diag}[\Lambda\{\widehat{D}_0, \widehat{D}(f)\} - D(f)],$$

and so that $\Gamma\{\widehat{D}_0, \widehat{D}(f)\}, \Lambda\{\widehat{D}_0, \widehat{D}(f)\}$ satisfy the last display of Proposition A.3. Hence, by definition of $\bar{M}_{\Omega^{-1}}$ we have

$$X_n = \bar{M}_{\Omega^{-1}} \begin{pmatrix} n^{1/2} \mathrm{vect}\left(\Gamma\{\widehat{D}_0, \widehat{D}(f)\} - I_p\right) \\ n^{1/2} \mathrm{diag}\left(\Lambda(\widehat{D}_0, \widehat{D}(f)) - D(f)\right) \end{pmatrix}.$$

Since $\bar{M}_{\Omega^{-1}}$ is fixed, we have, from Proposition A.3,

$$\begin{aligned} X_n &= \bar{M}_{\Omega^{-1}} E \begin{pmatrix} n^{1/2}\mathrm{vect}(\widehat{D}_0 - D_0) \\ n^{1/2}\mathrm{diag}\{\widehat{D}(f) - D(f)\} \end{pmatrix} + o_p(1) \\ &= \bar{M}_{\Omega^{-1}} E Y_n + o_p(1), \end{aligned}$$

say. Let $Q_{\mathrm{st},n}$ be the distribution of Y_n. Proposition A.2, with $\Omega = I_p$, $f_1(x) = \mathbf{1}_{\{x=0\}}$ and $f_2 = f$, implies as $n \to \infty$ that $d_w[Q_{\mathrm{st},n}, \mathcal{N}\{0, \tilde{V}(f)\}] \to 0$. Since $\bar{M}_{\Omega^{-1}}$, E and $\tilde{V}(f)$ are bounded as $n \to \infty$, this implies (by taking subsequences) that when $n \to \infty$,

$$d_w[Q_n, \mathcal{N}\{0, \bar{M}_{\Omega^{-1}} E \tilde{V}(f) E^{\mathrm{T}} \bar{M}_{\Omega^{-1}}^{\mathrm{T}}\}] \to 0.$$

Hence the proof is concluded. □

A.3 Asymptotics when diagonalizing more than two matrices

Proposition A.5. *Let $k \in \mathbb{N}$ be fixed. Let $f_1, \ldots, f_k : \mathbb{R}^d \to \mathbb{R}$ satisfy Condition A.3. Assume that there exists a fixed $\delta > 0$ so that for all $n \in \mathbb{N}$, for every pair $i \neq j$, $i, j = 1, \ldots, p$, there exists $l = 1, \ldots, k$ such that $|D(f_l)_{i,i} - D(f_l)_{j,j}| \geq \delta$. Let $\widehat{\Gamma} = \widehat{\Gamma}\{\widehat{D}_0, \widehat{D}(f_1), \ldots, \widehat{D}(f_k)\}$ be such that*

$$\widehat{\Gamma} \in \underset{\substack{\Gamma : \Gamma \widehat{D}_0 \Gamma^{\mathrm{T}} = I_p \\ \Gamma \text{ has rows } \gamma_1^{\mathrm{T}}, \ldots, \gamma_p^{\mathrm{T}}}}{\arg\max} \sum_{l=1}^{k} \sum_{j=1}^{p} \{\gamma_j^{\mathrm{T}} \widehat{D}(f_l) \gamma_j\}^2. \tag{6}$$

Then we can choose $\widehat{\Gamma}$ so that $\widehat{\Gamma} \to I_p$ in probability when $n \to \infty$.

Proof. Let, for U a $p \times p$ orthogonal matrix with rows $u_1^{\mathrm{T}}, \ldots, u_p^{\mathrm{T}}$,

$$\widehat{g}(U) = \sum_{l=1}^{k} \sum_{j=1}^{p} \{u_j^{\mathrm{T}} \widehat{D}_0^{-1/2} \widehat{D}(f_l) \widehat{D}_0^{-1/2} u_j\}^2.$$

Let

$$\begin{aligned} E_0 \quad = \quad & \{U \text{ orthogonal with rows } u_1^{\mathrm{T}}, \ldots, u_p^{\mathrm{T}}; \\ & \sum_{j=1}^{p} j(u_1)_j^2 \leq \cdots \leq \sum_{j=1}^{p} j(u_p)_j^2 \text{ and for } \ i = 1, \ldots, p, \sum_{j=1}^{p} U_{i,j} \geq 0\}. \end{aligned}$$

We observe that any orthogonal matrix can be obtained from a matrix in E_0, by row permutation and row multiplication by 1 or -1. Hence, for any n, there exists $\widehat{U}$ so that $\widehat{U} \in \arg\max_{U \in E_0} \widehat{g}(U)$ and $\widehat{U} \widehat{D}_0^{-1/2}$ satisfies (6).

We now aim at showing that $\widehat{U} \to I_p$ in probability as $n \to \infty$, which will conclude the proof since $\widehat{D}_0 \to I_p$ in probability. Assume that this is not the case. Then, there exists $\epsilon > 0$ and a subsequence $(n_m)_{m \in \mathbb{N}}$ so that for all $m \in \mathbb{N}$ and along n_m

$$\mathbb{P}(\|\widehat{U} - I_p\|_F \geq \epsilon) \geq \epsilon. \tag{7}$$

The matrices $D(f_1), \ldots, D(f_l)$ are bounded (this can be shown as in Proposition A.1). Hence, by compacity, up to extracting a further subsequence, we have that (7) holds along n_m and, as $m \to \infty$ and along n_m, $D(f_1) \to D_\infty(f_1), \ldots, D(f_k) \to D_\infty(f_k)$.

We let

$$g_\infty(U) = \sum_{l=1}^{k} \sum_{j=1}^{p} \{u_j^{\mathrm{T}} D_\infty(f_l) u_j\}^2.$$

We have, from Proposition A.1 and as observed in Miettinen et al. (2016), that, as $m \to \infty$ and along n_m,

$$\sup_{U \in U_0} |\widehat{g}(U) - g_\infty(U)| \to 0$$

in probability as $m \to \infty$. Hence, using a standard M-estimator argument and because E_0 is compact, if the unique maximum of g_∞ on E_0 is I_p, we obtain that, as $m \to \infty$ and along n_m, $\widehat{U} \to I_p$ in probability. This is contradictory to (7).

Hence, to conclude the proof, it suffices to show that the unique maximum of g_∞ on E_0 is I_p. We have

$$\begin{aligned} g_\infty(U) &= \sum_{l=1}^{k} [\|U^{\mathrm{T}} D_\infty(f_l) U\|_F^2 - \sum_{i \neq j} \{U^{\mathrm{T}} D_\infty(f_l) U\}_{i,j}^2] \qquad (8) \\ &\leq \sum_{l=1}^{k} \|U^{\mathrm{T}} D_\infty(f_l) U\|_F^2 \\ &= \sum_{l=1}^{k} \|D_\infty(f_l)\|_F^2. \end{aligned}$$

Also,

$$g_\infty(I_p) = \sum_{l=1}^{k} \sum_{j=1}^{p} D_\infty(f_l)_{j,j}^2 = \sum_{l=1}^{k} \|D_\infty(f_l)\|_F^2.$$

We next show that the identity matrix I_p is the unique maximizer of g_∞ in E_0. To see this, consider an arbitrary orthogonal matrix U which maximizes g_∞. From (8) we see that $U^{\mathrm{T}} D_\infty(f_l) U$ is a diagonal matrix for all $l = 1, \ldots, k$. Then, by its non-singularity, the matrix U must have a column with a non-zero first element. Call the first (from the left) such column of U by u. We show that all other elements of u must be zero. By the previous, u is an eigenvector of all $D_\infty(f_l)$ and we have,

$$D_\infty(f_l) u = \psi_l u, \quad \text{for all } l = 1, \ldots, k,$$

for some eigenvalues $\psi_l \in \mathbb{R}$, $l = 1, \ldots, k$. Assume then that u has a second non-zero element at some arbitrary position $q \neq 1$, meaning that both $u_1, u_q \neq 0$. Then we write

$$D_\infty(f_l)_{1,1} u_1 = \psi_l u_1 \quad \text{and} \quad D_\infty(f_l)_{q,q} u_q = \psi_l u_q, \quad \text{for all } l = 1, \ldots, k,$$

which in turn implies that $D_\infty(f_l)_{1,1} = D_\infty(f_l)_{q,q}$ for all $l = 1, \ldots, k$. By a continuity argument, this is a contradiction with the assumptions of the proposition. As the choice

of q was arbitrary, the only non-zero element in u is the first. Repeating now the same reasoning for other elements besides the first, we observe that each column of the maximizer U must have a single non-zero element, and by its orthogonality we have $U = PD$ for some permutation matrix P and some diagonal matrix D with diagonal components in $\{-1, 1\}$. The only matrix of that form belonging to E_0 is I_p and thus, for all $U \in E_0$ with $U \neq I_p$, we have $g(U) < g(I_p)$. □

Proposition A.6. *Assume the same setting and conditions as in Proposition A.5. For a diagonal matrix Λ, let $\Lambda_r = \Lambda_{r,r}$. Let $A_0, A_1, \ldots, A_k$ and B be $p^2 \times p^2$ matrices defined by*

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

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

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

and

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

Then, as $n \to \infty$, there exists $\widehat{\Gamma}$ satisfying (6) *so that*

$$n^{1/2}\mathrm{vect}(\widehat{\Gamma} - I_p) = B \begin{pmatrix} A_0 & A_1 & \ldots & A_k \end{pmatrix} \begin{pmatrix} n^{1/2}\mathrm{vect}(\widehat{D}_0 - D_0) \\ n^{1/2}\mathrm{vect}\{\widehat{D}(f_1) - D(f_1)\} \\ \vdots \\ n^{1/2}\mathrm{vect}\{\widehat{D}(f_k) - D(f_k)\} \end{pmatrix}.$$

Proof. Let $\widehat{\Gamma}$ satisfy (6) and $\widehat{\Gamma} \to I_p$ in probability when $n \to \infty$ (the existence follows from Proposition A.5). The proof of the proposition heavily follows the proofs of ii) in Theorem 2 of Miettinen et al. (2016) and Theorem 3 in Virta et al. (2018) and as such, we present below only some key steps.

From Proposition A.2 (taken in the case $\Omega = I_p$), we have $n^{1/2}(\widehat{D}_0 - I_p) = O_p(1)$ and $n^{1/2}\{\widehat{D}(f_l) - D(f_l)\} = O_p(1)$, for all $l = 1, \ldots, k$. By a continuity argument and our assumptions, we further have $D(f_1) \to D_\infty(f_1), \ldots, D(f_k) \to D_\infty(f_k)$ such that the limit matrices satisfy: there exists a fixed $\delta > 0$ so that for every pair $i \neq j$, $i, j = 1, \ldots, p$, there exists $l = 1, \ldots, k$ such that $|D_\infty(f_l)_{i,i} - D_\infty(f_l)_{j,j}| \geq \delta$. [The previous convergence holds up to extracting a subsequence. We omit this step in this proof for concision, but see the proof of Proposition A.3.] Finally, the rotation $\widehat{U}$ so that $\widehat{\Gamma} = \widehat{U}\widehat{D}_0^{-1/2}$ also satisfies $\widehat{U} \to I_p$ in probability.

Then, as in Virta et al. (2018), the maximization problem,

$$\underset{\substack{U:UU^{\mathrm{T}}=I_p \\ U \text{ has rows } u_1^{\mathrm{T}},\ldots,u_p^{\mathrm{T}}}}{\arg\max} \sum_{l=1}^{k}\sum_{j=1}^{p}\{u_j^{\mathrm{T}}\widehat{D}_0^{-1/2}\widehat{D}(f_l)\widehat{D}_0^{-1/2}u_j\}^2,$$

yields the estimation equations $\sqrt{n}\widehat{Y} = \sqrt{n}\widehat{Y}^{\mathrm{T}}$, where

$$\sqrt{n}\widehat{Y} = \sqrt{n}\sum_{l=1}^{k}\widehat{U}\widehat{R}(f_l)\widehat{U}^{\mathrm{T}}\mathrm{Diag}\{\widehat{U}\widehat{R}(f_l)\widehat{U}^{\mathrm{T}}\},$$

where we have used the shorthand $\widehat{R}(f_l) = \widehat{D}_0^{-1/2}\widehat{D}(f_l)\widehat{D}_0^{-1/2}$ and $\mathrm{Diag}(M)$ is equal to the square matrix M but with its off-diagonal elements set to zero. Linearizing the estimating equations asymptotically and vectorizing, we arrive at the following form,

$$\begin{aligned}&(I_p - K)\sum_{l=1}^{k}\Big\{[\mathrm{Diag}\{\widehat{U}D(f_l)\widehat{U}^{\mathrm{T}}\}\widehat{U}D(f_l)\otimes I_p] + [\mathrm{Diag}\{\widehat{U}D(f_l)\widehat{U}^{\mathrm{T}}\}\otimes D(f_l)]K\Big\}\\ &\cdot\sqrt{n}\mathrm{vec}(\widehat{U}-I_p) = -(I_p-K)\sqrt{n}\mathrm{vec}(\widehat{F}) + o_p(1),\end{aligned} \tag{9}$$

where K is the $p^2\times p^2$ commutation matrix satisfying $K^2 = I_p$, $\sqrt{n}\widehat{F} = \sum_{l=1}^{k}\sqrt{n}\{\widehat{R}(f_l) - D(f_l)\}D_\infty(f_l) = O_p(1)$ and $\mathrm{vec}(M) = (c_1^{\mathrm{T}},\ldots,c_k^{\mathrm{T}})^{\mathrm{T}}$ is the column vectorization where $c_1,\ldots,c_k$ are the k columns of a matrix M. The orthogonality constraint can be similarly linearized to yield,

$$\{(\widehat{U}\otimes I_p) + K\}\sqrt{n}\mathrm{vec}(\widehat{U}-I_p) = 0. \tag{10}$$

Summing (9) and (10), we obtain,

$$\widehat{A}\sqrt{n}\mathrm{vec}(\widehat{U}-I_p) = -(I_p-K)\sqrt{n}\mathrm{vec}(\widehat{F}) + o_p(1), \tag{11}$$

where $\widehat{A} \to A = (I_p-K)[\{\sum_{l=1}^{k}D_l^2\otimes I_p\} + \{\sum_{l=1}^{k}D_l\otimes D_l\}K] + I_p + K$ in probability, where we use the notation $D_l = D_\infty(f_l)$, $l = 1,\ldots,k$. Using the fact that $K(A\otimes B)K = B\otimes A$ for any conformable matrices A, B, we get the alternative form,

$$A = \left(\sum_{l=1}^{k}D_l^2\otimes I_p - \sum_{l=1}^{k}D_l\otimes D_l + I_p\right) + \left\{\sum_{l=1}^{k}D_l\otimes D_l - \sum_{l=1}^{k}I_p\otimes D_l^2 + I_p\right\}K.$$

Continuing as in Virta et al. (2018), each diagonal element of $\widehat{U}$ has a corresponding 1×1 diagonal block equal to 2 in A. Similarly, each pair of (a,b)th and (b,a)th off-diagonal elements in $\widehat{U}$ has a corresponding 2×2 sub-matrix in A of the form,

$$A_{ab} = \begin{pmatrix} 1+\sum_{l=1}^{k}d_{la}^2 - \sum_{l=1}^{k}d_{la}d_{lb} & 1-\sum_{l=1}^{k}d_{lb}^2 + \sum_{l=1}^{k}d_{la}d_{lb} \\ 1-\sum_{l=1}^{k}d_{la}^2 + \sum_{l=1}^{k}d_{la}d_{lb} & 1+\sum_{l=1}^{k}d_{lb}^2 - \sum_{l=1}^{k}d_{la}d_{lb}, \end{pmatrix},$$

where d_{la} is the ath diagonal element of D_l. The inverse of the sub-matrix is

$$A_{ab}^{-1} = \left\{ 2\sum_{l=1}^{k}(d_{la}-d_{lb})^2 \right\}^{-1} \begin{pmatrix} 1+\sum_{l=1}^{k} d_{lb}^2 - \sum_{l=1}^{k} d_{la}d_{lb} & \sum_{l=1}^{k} d_{lb}^2 - \sum_{l=1}^{k} d_{la}d_{lb} - 1 \\ \sum_{l=1}^{k} d_{la}^2 - \sum_{l=1}^{k} d_{la}d_{lb} - 1 & 1+\sum_{l=1}^{k} d_{la}^2 - \sum_{l=1}^{k} d_{la}d_{lb}, \end{pmatrix},$$

showing that A is invertible as by our assumptions $\sum_{l=1}^{k}(d_{la}-d_{lb})^2 \neq 0$ for all distinct pairs $a, b = 1, \ldots, p$. Thus, by Slutsky's theorem, we obtain from (11) that,

$$\sqrt{n}\mathrm{vec}(\widehat{U} - I_p) = -A^{-1}\sqrt{n}\mathrm{vec}(\widehat{F} - \widehat{F}^{\mathrm{T}}) + o_p(1),$$

showing that, $\sqrt{n}(\widehat{U} - I_p) = O_p(1)$. Consequently, also $\sqrt{n}(\widehat{\Gamma} - I_p) = O_p(1)$.

Finally, we next proceed as in the proof of ii) in Theorem 2 of Miettinen et al. (2016) to obtain that, as $n \to \infty$,

$$n^{1/2}(\widehat{\Gamma}_{i,i} - 1) = -(1/2)n^{1/2}(\widehat{D}_{0,i,i} - 1) + o_p(1)$$

and for $i \neq j$,

$$n^{1/2}\widehat{\Gamma}_{i,j} = \frac{\sum_{l=1}^{k}\{D(f_l)_i - D(f_l)_j\}n^{1/2}\{\widehat{D}(f_l)_{i,j} - D(f_l)_i\widehat{D}_{0,i,j}\}}{\sum_{r=1}^{k}\{D(f_r)_i - D(f_r)_j\}^2} + o_p(1).$$

Hence, the proposition follows from the definition of $A_0, A_1, \ldots, A_k, B$. □

Proposition A.7. *Assume the same settings and conditions as in Proposition A.5. Let $\Sigma(f, g)$ be as in Proposition A.2 with R replaced by $\mathrm{cov}(z)$ where z is the $np \times 1$ vector defined by, for $i = 1, \ldots, n$, $j = 1, \ldots, p$, $z_{(i-1)p+j} = Z_j(s_i)$. Let $f_0(x) = \mathbf{1}_{\{x=0\}}$. 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 sizes $p^2 \times p^2$ with block $(i+1), (j+1)$ equal to $\Sigma(f_i, f_j)$ for $i, j = 0, \ldots, p$.*

Let E be the $p^2 \times (k+1)p^2$ matrix defined by $E = B(A_0, A_1, \ldots A_k)$ (notation of Proposition A.6). Let $M_{\Omega^{-1}}$ be as in Proposition A.4 and let

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

Then, $\widehat{\Gamma} = \widehat{\Gamma}\{\widehat{M}_0, \widehat{M}(f_1), \ldots, \widehat{M}(f_k)\}$ satisfying (6) (with $\widehat{D}_0, \widehat{D}(f_1), \ldots, \widehat{D}(f_k)$ replaced by $\widehat{M}_0, \widehat{M}(f_1), \ldots, \widehat{M}(f_k)$) can be chosen so that, with Q_n the distribution of $n^{1/2}(\widehat{\Gamma} - \Omega^{-1})$, we have as $n \to \infty$

$$d_w\{Q_n, \mathcal{N}(0, F)\} \to 0.$$

Proof. The proof is the same as that of Proposition A.4. In particular, for $\widehat{\Gamma}\{\widehat{D}_0, \widehat{D}(f_1), \ldots, \widehat{D}(f_k)\}$ satisfying (6), the matrix $\widehat{\Gamma}\{\widehat{D}_0, \widehat{D}(f_1), \ldots, \widehat{D}(f_k)\}\Omega^{-1}$ satisfies (6) (with $\widehat{D}_0, \widehat{D}(f_1), \ldots, \widehat{D}(f_k)$ replaced by $\widehat{M}_0, \widehat{M}(f_1), \ldots, \widehat{M}(f_k)$). □

Proposition A.8. *Assume the same settings and conditions as in Proposition A.5. Let $(\widehat{\Gamma}_n)_{n\in\mathbb{N}}$ be any sequence of $p \times p$ matrices so that for any $n \in \mathbb{N}$, $\widehat{\Gamma}_n = \widehat{\Gamma}_n\{\widehat{M}_0, \widehat{M}(f_1), \ldots, \widehat{M}(f_k)\}$ satisfies (6) (with $\widehat{D}_0, \widehat{D}(f_1), \ldots, \widehat{D}(f_k)$ replaced by $\widehat{M}_0, \widehat{M}(f_1), \ldots, \widehat{M}(f_k)$). Then, there exists a sequence of permutation matrices (P_n) and a sequence of diagonal matrices (D_n), with diagonal components in $\{-1, 1\}$ so that, with $\check{\Gamma}_n = D_nP_n\widehat{\Gamma}_n$, the sequence $(\check{\Gamma}_n)$ satisfies the conclusions of Proposition A.5, with the limit I_p replaced by Ω^{-1}, and of Proposition A.7.*

Proof. With the notation of the proof of Proposition A.5, for $\widehat{\Gamma}_n$ satisfying (6), there exists P_n, D_n, as described in the proposition, so that $D_n P_n \widehat{\Gamma}_n \widehat{D}_0^{1/2} \in E_0$ and $D_n P_n \widehat{\Gamma}_n$ satisfies (6). Hence, with the same argument as in the proof of the last part of Proposition A.5, we have $D_n P_n \widehat{\Gamma}_n \to I_p$ in probability as $n \to \infty$. Furthermore, as in the proof of Proposition A.7, we can show that $D_n P_n \widehat{\Gamma}_n$ satisfies the conclusion of this proposition. Finally, the proof is concluded by observing that any matrix $\bar{\Gamma}$ satisfies (6) (with $\widehat{D}_0, \widehat{D}(f_1), \ldots, \widehat{D}(f_k)$ replaced by $\widehat{M}_0, \widehat{M}(f_1), \ldots, \widehat{M}(f_k)$) if and only if the corresponding matrix $\bar{\Gamma}\Omega$ satisfies (6). □

The results of Propositions 3 and 4 derive directly from Proposition A.8.

Lemma A.2. *Let Conditions A.1 and A.2 hold. Let f satisfy Condition A.3. Let $\bar{X} = n^{-1} \sum_{i=1}^n X(s_i)$. Let*

$$\widehat{M}_{\mathrm{st}}(f) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n f(s_i - s_j)\{X(s_i) - \bar{X}\}\{X(s_j) - \bar{X}\}^{\mathrm{T}}.$$

Then as $n \to \infty$

$$\widehat{M}_{\mathrm{st}}(f) - \widehat{M}(f) = O_p\left(n^{-1}\right).$$

Proof. Let $k, l \in \{1, \ldots, p\}$ and let $f_{i,j} = f(s_i - s_j)$. We have

$$\begin{aligned}
\widehat{M}_{\mathrm{st}}(f)_{k,l} - \widehat{M}(f)_{k,l} &= n^{-1} \sum_{i=1}^n \sum_{j=1}^n f_{i,j}\{X_k(s_i)X_l(s_j) - X_k(s_i)\bar{X}_l - \bar{X}_k X_l(s_j) + \bar{X}_k \bar{X}_l\} \\
&\quad - n^{-1} \sum_{i=1}^n \sum_{j=1}^n f_{i,j} X_k(s_i) X_l(s_j) \\
&= -\bar{X}_l\{n^{-1} \sum_{i=1}^n \sum_{j=1}^n f_{i,j} X_k(s_i)\} - \bar{X}_k\{n^{-1} \sum_{i=1}^n \sum_{j=1}^n f_{i,j} X_l(s_j)\} \\
&\quad + n^{-1} \sum_{i=1}^n \sum_{j=1}^n f_{i,j} \bar{X}_k \bar{X}_l. \qquad (12)
\end{aligned}$$

Now, for $q = 1, \ldots, p$, $E(\bar{X}_q) = 0$ and

$$\mathrm{var}(\bar{X}_q) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \mathrm{cov}\{X_q(s_i), X_q(s_j)\}.$$

Also, $\max_{i=1,\ldots,n} \sum_{j=1}^n |\mathrm{cov}\{X_q(s_i), X_q(s_j)\}|$ is bounded because of (2) and Lemma 4 in Furrer et al. (2016). Hence $\mathrm{var}(\bar{X}_q) = O(1/n)$ and so $\bar{X}_q = O_p(n^{-1/2})$.

Also, let

$$\epsilon_q = n^{-1} \sum_{i=1}^n \sum_{j=1}^n f_{i,j} X_q(s_i).$$

Then $E(\epsilon_q) = 0$ and

$$\mathrm{var}(\epsilon_q) = n^{-2} \sum_{i=1}^{n} \sum_{k=1}^{n} (\sum_{j=1}^{n} f_{i,j})(\sum_{j=1}^{n} f_{k,j}) \mathrm{cov}\{X_q(s_i), X_q(s_k)\}.$$

From Condition A.3 and Lemma 4 in Furrer et al. (2016), there exists a finite constant H so that

$$\max_{i=1,\ldots,n} \sum_{j=1}^{n} |f_{i,j}| \leq H.$$

Hence

$$\begin{aligned} \mathrm{var}(\epsilon_q) &\leq H^2 n^{-2} \sum_{i=1}^{n} \sum_{k=1}^{n} |\mathrm{cov}\{X_q(s_i), X_q(s_k)\}| \\ &= O\left(n^{-1}\right) \end{aligned}$$

as before. Hence $\epsilon_q = O_p(n^{-1/2})$. Also, we have seen above that

$$n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} |f_{i,j}|$$

is bounded. Hence, from (12), we conclude the proof of the lemma. □

References

Bachoc, F. (2014). Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes. *Journal of Multivariate Analysis*, 125:1–35.

Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., and Moulines, E. (1997). A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444.

Comon, P. and Jutten, C. (2010). *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press.

Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons, Inc., New York, 2nd edition.

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Eddelbuettel, D. and Sanderson, C. (2014). Rcpparmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063.

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

Filzmoser, P. (2015). *StatDA: Statistical Analysis for Environmental Data*. R package version 1.6.9.

Furrer, R., Bachoc, F., and Du, J. (2016). Asymptotic properties of multivariate tapering for estimation and prediction. *Journal of Multivariate Analysis*, 149:177–191.

Furrer, R. and Genton, M. G. (2011). Aggregation-cokriging for highly multivariate spatial data. *Biometrika*, 98:615–631.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST*, 13:263–312.

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

Goulard, M. and Voltz, M. (1992). Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix. *Mathematical Geology*, 24(3):269–286.

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

Ilmonen, P., Nevalainen, J., and Oja, H. (2010a). Characteristics of multivariate distributions and the invariant coordinate system. *Statistics & Probability Letters*, 80(23):1844–1853.

Ilmonen, P., Nordhausen, K., Oja, H., and Ollila, E. (2010b). A new performance index for ICA: properties, computation and asymptotic analysis. In *Latent Variable Analysis and Signal Separation*, pages 229–236. Springer.

Istas, J. and Lang, G. (1997). Quadratic variations and estimation of the local Hölder index of a gaussian process. *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, 33(4):407–436.

Matilainen, M., Nordhausen, K., and Oja, H. (2015). New independent component analysis tools for time series. *Statistics & Probability Letters*, 105:80–87.

Miettinen, J., Illner, K., Nordhausen, K., Oja, H., Taskinen, S., and Theis, F. (2016). Separation of uncorrelated stationary time series using autocovariance matrices. *Journal of Time Series Analysis*, 37(3):337–354.

Miettinen, J., Nordhausen, K., Oja, H., and Taskinen, S. (2012). Statistical properties of a blind source separation estimator for stationary time series. *Statistics & Probability Letters*, 82:1865–1873.

Miettinen, J., Nordhausen, K., Oja, H., and Taskinen, S. (2014). Deflation-based separation of uncorrelated stationary time series. *Journal of Multivariate Analysis*, 123:214–227.

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

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

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

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reimann, C., Filzmoser, P., Garrett, R., and Dutter, R. (2008). *Statistical Data Analysis Explained. Applied Environmental Statistics with R*. Wiley, Chichester.

Ribeiro Jr, P. J. and Diggle, P. J. (2016). *geoR: Analysis of Geostatistical Data*. R package version 1.7-5.2.

Sun, D. and Sun, J. (2002). Strong semismoothness of eigenvalues of symmetric matrices and its application to inverse eigenvalue problems. *SIAM Journal on Numerical Analysis*, 40(6):2352–2367.

Virta, J., Lietzén, N., Ilmonen, P., and Nordhausen, K. (2018). Asymptotically and computationally efficient tensorial JADE. *arXiv preprint arXiv:1808.00791*.

Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*. Springer-Verlag, Berlin, 3rd edition.

Xie, T. and Myers, D. E. (1995). Fitting matrix-valued variogram models by simultaneous diagonalization (Part I: Theory). *Mathematical Geology*, 27:867–875.

Xie, T., Myers, D. E., and Long, A. E. (1995). Fitting matrix-valued variogram models by simultaneous diagonalization (Part II: Application). *Mathematical Geology*, 27:877–888.

Zhang, H. (2007). Maximum-likelihood estimation for multivariate spatial linear coregionalization models. *Environmetrics*, 18:125–139.